Lang Resources & Evaluation (2022) 56:165–189
https://doi.org/10.1007/s10579-021-09544-6

ORIGINAL PAPER

A large and evolving cognate database

Khuyagbaatar Batsuren[1] · **Gábor Bella**[2] · **Fausto Giunchiglia**[2,3]

Accepted: 13 May 2021 / Published online: 30 May 2021

Abstract We present *CogNet*, a large-scale, automatically-built database of sense-tagged *cognates*—words of common origin and meaning across languages. CogNet is continuously evolving: its current version contains over 8 million cognate pairs over 338 languages and 35 writing systems, with new releases already in preparation. The paper presents the algorithm and input resources used for its computation, an evaluation of the result, as well as a quantitative analysis of cognate data leading to novel insights on language diversity. Furthermore, as an example on the use of large-scale cross-lingual knowledge bases for improving the quality of multilingual applications, we present a case study on the use of CogNet for bilingual lexicon induction in the framework of cross-lingual transfer learning.

Keywords Cognate · Lexical semantics · Lexical database

✉ Gábor Bella
gabor.bella@unitn.it

Khuyagbaatar Batsuren
khuyagbaatar@num.edu.mn

Fausto Giunchiglia
fausto.giunchiglia@unitn.it

[1] Department of Information and Computer Science, National University of Mongolia, Ikh surguuliin gudamj 1, 14200 Ulaanbaatar, Mongolia

[2] Department of Information Engineering and Computer Science, University of Trento, via Sommarive 5, 38123 Trento, Italy

[3] College of Computer Science and Technology, Jilin University, Changchun, China

1 Introduction

Cognates are words in different languages that share a common origin and the same meaning, such as the English *letter* and the French *lettre*. Cognates have been traditionally studied in various fields of linguistics, such as historical linguistics (Crowley & Bowern, 2010; List, 2019b) or in the context of the study of language diversity (Giunchiglia et al., 2017). More recently, there has been a surge in the use of structured lexical knowledge in order to improve the performance of cross-lingual language processing tasks such as machine translation, lexicon extension and induction, or bilingual word embedding (Artetxe et al., 2016). The use of knowledge in such tasks has been on the rise especially due to an increasing recognition of the inherent limits of purely corpus-driven approaches.

However, despite the specific use of cognate data in several works (Kondrak et al., 2003; Tsvetkov & Dyer, 2015; Wu & Yarowsky, 2018), the improvements obtained are marginal due to the typically low coverage of existing cognate databases. Furthermore, existing databases offer limited practical benefit from an applied perspective, for two reasons. Firstly, popular databases that are used by cognacy-based methods in historical linguistics, such as ASJP (Jäger, 2018; Wichmann et al., 2010), IELex (Bouckaert et al., 2012), or ABVD (Greenhill et al., 2008), have by design a low lexical coverage of typically less than a hundred basic concepts per language, but with an extremely broad coverage of up to 4000 languages. Secondly, in these databases, lexical entries that belong to scripts other than Latin or Cyrillic mostly appear in phonetic transcription instead of their actual orthographies in their original scripts, limiting their use for processing written text.

This paper describes *CogNet*, a large-scale multilingual cognate database. CogNet is well adapted to the needs of computational applications over written language, such as machine translation or other cross-lingual corpus-based techniques. Firstly, its coverage is very high: it contains 8.1 million sense-tagged word pairs in 338 languages and 35 writing systems. Secondly, its precision is evaluated to 95%, high enough for an efficient use in computational tasks. Thirdly, it contains words in their original orthographies (as opposed to phonetic transcription), allowing their direct reuse for the processing of written text. And lastly, it is in constant evolution, with improving precision and recall as its input resources are themselves evolving.

The main technical contributions of this article are:

1. a generic method that uses a set of resources as *sources of evidence* for inferring cognate relationships across words;
2. CogNet, a freely downloadable, large-scale cognate database generated with the method above, with each cognate pair labelled according to the evidence on which it is based (strictly etymological or also indirect evidence);
3. the extraction of massive amounts of linguistic knowledge—phonetic transcriptions, transliteration rules, etymological relationships—from Wiktionary, in order to provide evidence to the algorithm;

4. an analysis of cognate data with respect to *language diversity*, i.e. the universality or locality of cognates, providing insights both from theoretical and computational viewpoints;
5. an example study on the use of CogNet data to improve state-of-the-art results on the well-known task of bilingual lexicon induction.

With respect to an initial release of CogNet, introduced in (Batsuren et al., 2019a), this paper presents a method and a resource that have been greatly extended and redesigned for extensibility by new input resources. Consequently, the size of the CogNet database has been multiplied by 2.5 with respect to its first version. The algorithm has been extended to take new forms of linguistic evidence into account. At the same time, precision was improved above 95%, according to new evaluations over a much larger and more diverse set of gold standard databases. This paper also contains a discussion of the use of CogNet data for the study of linguistic diversity, as well a section dedicated to using CogNet as seed data in a task of bilingual lexicon induction through cross-lingual transfer learning, currently a heavily researched topic in computational linguistics.

All CogNet versions are made available on the CogNet website.[1] The cognate data has also been published in a browseable form on the website of the Universal Knowledge Core multilingual database[2] for the visual exploration of cognate data. In the long term, this web tool is intended for linguists both for the exploration of data and for collaborative work on multilingual language resources.

The paper is organised as follows. Section 2 presents the state of the art. Section 3 provides the main principles and the high-level architecture of our solution for computing CogNet. Section 4 describes the main cognate computation algorithm. Sections 5 to 8 present the different sources of evidence of cognacy (semantic, etymological, phonetic, orthographic, geographic), the way they are exploited by the algorithm, and the way they can be extended in order to improve CogNet. The method is parametrised, and then evaluated in Sect. 9 using three distinct evaluation methods. Section 10 presents the resulting CogNet database and its successive evolving versions. Section 11 presents initial analytics results over cognate data, providing insightful observations on the relatedness and the diversity of languages. Section 12 provides an example study on the use of CogNet data for bilingual lexicon induction. Finally, Sect. 13 concludes the paper.

2 State of the art

Cognates have so far been defined and explored in two fundamental ways by two distinct communities in linguistics. On the one hand, *cognate identification* has been studied within linguistic typology and historical linguistics. On the other hand, computational linguists have been researching methods for *cognate production* and,

[1] http://cognet.ukc.disi.unitn.it.

[2] http://ukc.datascientia.eu.

to some extent, have tried to use cognate data for cross-lingual applications such as machine translation and bilingual lexicon induction.

To these two communities correspond two considerably different definitions of cognacy. In historical linguistics, where the historical relatedness of languages is studied, cognates must have a provable etymological relationship and must be fully absorbed into each language. Accordingly, the English *father* and the French *père* are considered as cognates due to a common ancestor, but the much more similar English *song* and the Japanese ソング (/songu/) are not, the latter being considered as a loanword. In computational linguistics, the notion of cognate is more relaxed with respect to etymology and loanwords are also considered as cognates (Kondrak et al., 2003). The methods presented in this paper are inscribed in a computational context which is better suited by the more relaxed interpretation of cognacy. Still, by annotating our output cognates by the kind of linguistic evidence on which they are based (purely etymological or also others), we leave the possibility of exploiting CogNet to other research communities as well.

In historical linguistics, cognate identification methods are mainly based on three types of similarity measures: semantic, phonetic, and orthographic. For information on semantic similarity, special-purpose multilingual dictionaries, such as the well-known *Swadesh List*, are used, and more recently cross-lingual word embeddings (Kanojia et al., 2020, 2021). For orthographic similarity, string metrics (Hauer & Kondrak, 2011; St Arnaud et al., 2017) are often employed, e.g. edit distance, Dice's coefficient, or LCSR. As these methods do not work across scripts, they are completed by phonetic similarity, exploiting transformations and sound changes across related languages (Kondrak, 2000; List, 2012; Jäger, 2013; Jäger et al., 2017; Rama et al., 2017). Phonetic similarity measures, however, require phonetic transcriptions to be *a priori* available. More recently, historical linguists have started exploiting identified cognates to infer phylogenetic relationships across languages (Rama et al., 2018; Jäger, 2018), and phylogenetic approaches to identify lexical borrowings (List, 2019a).

In computational linguistics, cognate production consists of finding for a word in a given language its cognate pair in another language. State-of-the-art methods (Beinborn et al., 2013; Sennrich et al., 2016; Ciobanu & Dinu, 2020) have employed character-based machine translation, trained from parallel corpora, to produce cognates or transliterations. (Wu & Yarowsky, 2018) also employ similar techniques, as well as multilingual dictionaries, to produce large-scale cognate clusters for Romance and Turkic languages. Although the cognates produced in this manner are, in principle, a good source for improving certain cross-lingual tasks in NLP, the quality of these methods often suffers due to not being able to handle certain linguistic phenomena properly. For example, words in languages such as Arabic or Hebrew are written without vowels and machine-produced transliterations often fail to vowelize such words (Karimi et al., 2011). The solution we propose is the use of a dictionary-based transliteration tool over machine transliteration.

Our method provides new contributions with respect to existing results in both fields. Firstly, to our knowledge no other work on cognate generation has so far used high-quality multilingual lexical resources on a scale as large as ours, covering hundreds of languages and more than 100,000 cross-lingual concepts. Secondly,

such a large coverage could only be achieved thanks to a robust transliteration tool that is part of our contributions. Finally, our novel, combined use of multiple—phonetic, orthographic, semantic, geographic, and etymological—sources of evidence for detecting cognates was crucial to obtain high-quality results, in terms of both precision and recall.

3 Principles and architecture

The level of correctness of any linguistic resource has a major influence on downstream tasks exploiting it. For this reason, an important design choice for our method is to favour precision over recall. We have consequently adopted a definition of cognacy that is more strict than conventional definitions used in computational linguistics (Kondrak et al., 2003): *two words in different languages are cognates if they have the same meaning and there is sufficient evidence of their etymological relatedness (common ancestry or borrowing).*

Based on this interpretation, our method for deciding on whether there is a cognate relationship between a pair of words is based on three main sources of evidence:

(a) *evidence of semantic equivalence*, i.e. that the two words share at least one common meaning;
(b) *evidence of etymological relatedness*; and
(c) the *logical transitivity* of the cognate relationship.

These three principles are applied in three distinct steps in our method, as shown in Fig. 1. Firstly, all candidate cross-lingual word pairs (i.e. coming from different languages) that fulfil the criterion (a) of sharing a common meaning are retrieved from a *multilingual lexical database*. Secondly, the *CogNet algorithm* emits a decision of whether the word pair in input are cognates, based on a set of resources providing direct or indirect etymological evidence according to criterion (b). And thirdly, before collecting the resulting cognate relationships in the CogNet database, evidence by logical transitivity from criterion (c) is applied.

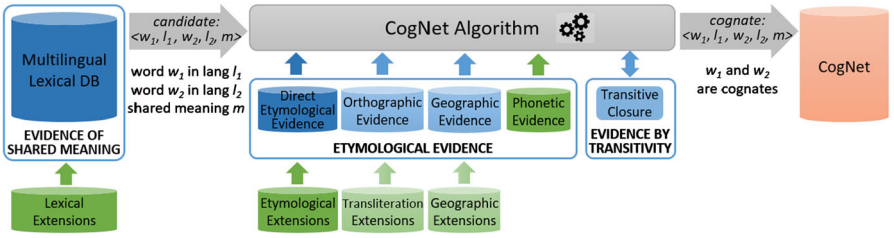

Fig. 1 The architecture of the cognate computation method, including the underlying extensible resources. The colour represents the CogNet version where the resource was added (dark blue: CogNet v0, light blue: v1, dark green: v2, light green: future v3 release)

Cross-lingual word pairs with evidence of shared meaning can be retrieved from multilingual lexico-semantic databases that provide equivalent meanings across languages. Examples of such databases are the *Open Multilingual WordNet* (Bond & Foster, 2013), *BabelNet* (Navigli & Ponzetto, 2010), or the *Universal Knowledge Core* (Giunchiglia et al., 2018) that we used in our research.

For the CogNet algorithm we consider two kinds of etymological evidence: *direct* and *indirect*. Direct evidence is provided by gold-standard etymological resources, such as the *Etymological WordNet* that we present in Sect. 6. Such evidence, however, is relatively sparse and would provide low recall. We therefore also consider indirect evidence: the geographic proximity of language speakers, combined with the phonetic or orthographic similarity of words, can provide strong clues on language contact and probable cross-lingual lexical borrowing. For each cognate pair predicted, we keep track of the kind of evidence (direct or indirect) it is based on, and annotate the output accordingly. This allows CogNet to be used with respect to different requirements of cognacy, such as computational linguistics that relies on a more relaxed interpretation and accepts, e.g., loanwords as cognates, or historical linguistics that relies on etymology alone—provided that the input evidence of CogNet itself is deemed reliable by the scientific community.

Finally, *evidence by logical transitivity* means that if words w_a and w_b are cognates and w_b and w_c are cognates then w_a and w_c are also cognates. For example, if the German *Katze* is found to be a cognate (via direct etymological evidence) of the English *cat* and *cat* is found to be a cognate (via indirect evidence) of the French *chat* then *Katze* and *chat* are also considered to be cognates.

The architecture of our method is designed to be extensible: while the CogNet algorithm relies on external sources of evidence, its logic itself is resource-agnostic. Extension serves the improvement of the recall and/or the precision of the output. As also shown in Fig. 1, there are two fundamental ways CogNet can be extended (extensions are represented as resources in light or dark green): either by increasing the size of an existing source of evidence or by introducing new kinds of evidence (such as phonetic evidence in Fig. 1). We took advantage of both approaches in order massively to increase recall within each version of CogNet, as explained in Sect. 10.

4 The CogNet algorithm

The centrepiece of the architecture in Fig. 1 is the CogNet algorithm, shown in p. 6. Its input is a single lexical concept from the multilingual lexical DB, the algorithm being applicable to every concept in loop. It builds multiple versions of the undirected CogNet graph *CogNet* where each node represents a word and each edge between two nodes represents a cognate relationship. The multiple versions correspond to using different subsets of evidence, and thus to different definitions of cognacy.

Algorithm 1: The CogNet Algorithm

Input : m, a lexical concept (shared meaning)
Input : $\mathcal{DB}$, a multilingual lexical database
Output: three cognate graphs with the meaning m:
$CogNet_{\text{DIR}}$, based on direct evidence only,
$CogNet^{+}_{\text{DIR}}$, based on direct evidence plus transitivity, and
$CogNet^{+}_{\text{IND}}$, based on both direct and indirect evidence, as well as transitivity

1 $V, E_{\text{DIR}}, E_{\text{IND}} \leftarrow \emptyset$;
2 $\mathcal{L} \leftarrow \text{Languages}_{\mathcal{DB}}(m)$;
3 **for** *each language* $l \in \mathcal{L}$ **do**
4 **for** *each word* $w \in \textit{Words}_{\mathcal{DB}}(m, l)$ **do**
5 $V \leftarrow V \cup \{v = <w, l>\}$;
6 **for** *each node* $v_1 = <w_1, l_1> \in V$ **do**
7 **for** *each node* $v_2 = <w_2, l_2> \in V$ **do**
8 **if** $l_1 = l_2$ **then**
9 continue;
10 **if** *DirectEvidence*(w_1, l_1, w_2, l_2) **then**
11 $E_{\text{DIR}} \leftarrow E_{\text{DIR}} \cup \{e = <v_1, v_2>\}$;
12 **else if** *PhonSim*$(w_1, l_1, w_2, l_2)) + T_G \times GeoProx(l_1, l_2) > T_F$ **then**
13 $E_{\text{IND}} \leftarrow E_{\text{IND}} \cup \{e = <v_1, v_2>\}$;
14 **else if** *OrthSim*$(w_1, l_1, w_2, l_2)) + T_G \times GeoProx(l_1, l_2) > T_F$ **then**
15 $E_{\text{IND}} \leftarrow E_{\text{IND}} \cup \{e = <v_1, v_2>\}$;
16 $CogNet_{\text{DIR}} \leftarrow <\text{V}, E_{\text{DIR}}>$;
17 $CogNet_{\text{IND}} \leftarrow <\text{V}, E_{\text{DIR}} \cup E_{\text{IND}}>$;
18 $CogNet^{+}_{\text{DIR}} = \text{TransitiveClosure}(CogNet_{\text{DIR}})$
19 $CogNet^{+}_{\text{IND}} = \text{TransitiveClosure}(CogNet_{\text{IND}})$
20 **return** $<CogNet_{\text{DIR}}, CogNet^{+}_{\text{DIR}}, CogNet^{+}_{\text{IND}}>$;

As shown in algorithm 1, p. 6, the process starts by retrieving the lexicalisations of the input concept in all available languages and creating the corresponding word nodes in the graph (lines 2–5). All such words thus fulfil the criterion of semantic equivalence above. Then, for all different-language word pairs that express the concept (lines 6–9), we verify whether etymological evidence exists for a potential cognate relationship. The latter may either be direct (line 10) or indirect evidence (lines 12–15). Indirect evidence is computed as a score of relatedness combined of phonetic similarity (*PhonSim*), orthographic similarity (*OrthSim*), and geographic proximity (*GeoProx*). We consider indirect evidence to be sufficient if a combined score is superior to a threshold T_F, computed together with T_G as hyperparameters of the algorithm. For the two kinds of evidence, two kinds of edges are created between the word nodes (lines 11, 13, 15), collected in E_{DIR} and E_{IND}. As the last step, in order to apply the principle of logical transitivity, the *transitive closure* of the graph is computed separately for direct evidence (lines 16, 18) and for combined direct and indirect evidence (lines 17, 19).

In the three graphs returned as output $CogNet_{\mathrm{DIR}}$, $CogNet^{+}_{\mathrm{DIR}}$, and $CogNet^{+}_{\mathrm{IND}}$, each connected subgraph represents a group of cognate words. The last one, $CogNet^{+}_{\mathrm{IND}}$, is the full set of cognates based on all kinds of evidence available to the algorithm. The first one, $CogNet_{\mathrm{DIR}}$, is a subset based solely on equivalence of word meaning and direct etymological evidence, excluding loanwords. Finally, $CogNet^{+}_{\mathrm{DIR}}$ extends this subset by transitive closure over cognates that express the same meaning.

5 Evidence of shared meaning

The CogNet algorithm takes as input word pairs that fulfil the following criteria: they are from different languages and they have the same meaning. Bilingual and multilingual dictionaries are able to provide such information. In order to maximise recall, the most effective choice are large multilingual lexical databases that cover hundreds of languages and millions of words, built from the aggregation of smaller resources. Examples of such databases are *PanLex* (Kamholz et al., 2014), *BabelNet*, or the *Universal Knowledge Core* (UKC). Beyond coverage, however, the quality and linguistic relevance of input resources also needs to be taken into consideration. Thus, automatically built resources not validated by human experts or language speakers run the risk of introducing language mistakes that are then propagated into CogNet. Likewise, resources with a coverage markedly different from the general vocabulary of a language (e.g. encyclopaedic) may introduce a non-linguistic bias into the results.

The current, second version of CogNet relies on two multilingual lexical databases: the UKC and PanLex. The UKC, already used in linguistics research and practical applications (Bella et al., 2016; Giunchiglia et al., 2017; Bella et al., 2017), includes the lexicons and lexico-semantic relations for over 1100 languages, containing over 2 million words and over 3 million language-specific word meanings.[3] By design, it concentrates on the common vocabulary and excludes proper nouns and named entities in general. It was built from high-quality, human-curated *wordnets* (Miller, 1995) and *wiktionaries* converted into wordnets (Bond & Foster, 2013). As most wordnets map their units of meaning (*synsets* in WordNet terminology) to English meanings, they can effectively be interconnected into a cross-lingual lexical resource. The UKC reifies all of these mappings as supra-lingual *lexical concepts* (over 110k in total). For example, if the German *Fahrrad* and the Italian *bicicletta* are mapped to the English *bicycle* then a single concept is created to which all three language-specific meanings (i.e., wordnet synsets) will be mapped.

In order further to extend the lexical coverage of the UKC, CogNet also relies on about 500,000 additional words retrieved from the PanLex lexical database and aligned in meaning with the UKC concepts, bringing the size of its input vocabulary to about 2.5 million words.

[3] http://ukc.datascientia.eu

6 Direct etymological evidence

By *direct etymological evidence* we understand a database of human-validated etymological ancestor–descendant relationships between words. Algorithm 1 exploits *etymological ancestor* (marked as *Anc* below) relations for each word of the word pair being evaluated as cognates. Two words are considered as etymologically related if they are found to have at least one common etymological ancestor word (such as the German *Ross* and the English *horse* having as ancestor the proto-Germanic root **harss-*).

$$\text{DirectEvidence}(w_1, l_1, w_2, l_2) = \begin{cases} \text{true} & \text{if } \text{Anc}(w_1, l_1) \cap \text{Anc}(w_2, l_2) \neq \emptyset \\ \text{false} & \text{otherwise} \end{cases} \quad (1)$$

CogNet retrieves ancestor relations from the *Etymological WordNet* (EWN)[4] (De Melo, 2014), as well as Wiktionary. EWN is a lexical resource that provides relations between words, e.g. derivational or etymological. EWN was automatically built by harvesting etymological information encoded in the 2013 version of Wiktionary, and has provided 94,832 cross-lingual etymological relations to CogNet. While this corresponds to only about 3% of all cognate pairs in CogNet, the overall contribution of direct etymological evidence is much higher due to the transitivity of the cognate relationship.

Wiktionary contains a much larger and still growing number of etymological entries, which we also reused in CogNet in order to extend the size of our etymological evidence. Thus, we increased the 95k etymological relations from EWN by six-fold, to a total of 673k relations,[5] which contributed to the significant increase of precision and recall in the last version of CogNet. Section 9 provides evaluation results on the increase in precision and recall due to the extension of etymological data.

7 Phonetic and orthographic evidence

Phonetic evidence has been a major device for research in historical linguistics, the idea being that similar pronunciation, in combination with other sources of evidence such as equivalent meanings, is a likely proof of a common etymological origin. CogNet is able to increase its recall massively because, in case of missing phonetic evidence, it considers similar orthography as indirect evidence for a phonetic resemblance.

CogNet relies on phonetic and orthographic evidence alike. Both are computed using a string similarity metric *LCSSim* based on the *longest common subsequence* (LCS) of the two input words, returning a similarity score between 0 and 1:

[4] http://www1.icsi.berkeley.edu/~demelo/etymwn/.

[5] https://github.com/kbatsuren/CogNet/tree/master/supporting%20resources.

$$\mathrm{LCSSim}(w_1, w_2) = \frac{2 \times \mathrm{len}(\mathrm{LCS}(w_1, w_2))}{\mathrm{len}(w_1) + \mathrm{len}(w_2)} \tag{2}$$

Phonetic similarity is based on the phonetic transcription (*PhonTra*) of words:

$$\mathrm{PhonSim}(w_1, w_2) = \mathrm{LCSSim}(\mathrm{PhonTra}(w_1), \mathrm{PhonTra}(w_2)) \tag{3}$$

We only used human-provided, and thus high-reliability phonetic transcriptions that we extracted from Wiktionary, wherever available. Since the IPA phonetic transcriptions in Wiktionary are complex (107 letters, 52 diacritics, and 4 prosodic marks) and thus cannot efficiently be fed to *LCSSim*, we converted them to the simpler ASJP representation (41 letters) using the *asjp* Python tool.[6] This way we obtained phonetic transcriptions for over 443k words.[7]

An improvement of phonetic similarity computation we foresee for future versions of CogNet is to replace our *PhonSim* function by the use of *Pointwise Mutual Information* data between words using ASJP transcriptions, as proposed by Jäger (2018).

CogNet uses orthographic similarity to compare words when their phonetic transcriptions are not available. When the input words w_1 and w_2 are written in the same script, they are directly compared using *LCSSim*. On the other hand, when they belong to different writing systems, LCS returns 0 and thus the formula above is not directly usable. In order to be able to identify cognates across writing systems, we apply transliteration to the Latin script (also known as *romanisation*) using our own *WikTra* tool. Orthographic similarity is thus computed as:

$$\mathrm{OrthSim}(w_1, w_2) = \max\{\mathrm{LCSSim}(w_1, w_2), \mathrm{LCSSim}(\mathrm{WikTra}(w_1), \mathrm{WikTra}(w_2))\} \tag{4}$$

WikTra is a dictionary-based transliteration tool compiled from information collected from *Wiktionary* and developed specifically for this work by the authors.[8] It is Unicode-based and supports 85 languages in 35 writing systems, defining transliteration rules and codes according to international standards, as developed by the Wiktionary community (the largest community in lexicography).

The use of *WikTra* as opposed to other existing tools is justified by a need for high-quality results that also cover complex cases of orthography, e.g. in Semitic scripts where vowels are typically omitted. In particular, *Junidecode*[9] is a character-based transliterator, an approach that seriously limits its accuracy. The *Google transliterator* is dictionary-based and is therefore of higher quality, but it supports a lower number of languages and is not freely available. *Uroman* (Hermjakob et al., 2018) is a new, high-quality, dictionary-based tool that nevertheless provides a limited support for scripts without vowels (e.g., Arabic or Hebrew).

[6] https://pypi.org/project/asjp/.

[7] https://github.com/kbatsuren/CogNet/tree/master/supporting%20resources.

[8] https://github.com/kbatsuren/wiktra.

[9] http://github.com/gcardone/junidecode.

WikTra still needs to be improved for Thai and Japanese. In Thai, WikTra only works on monosyllabic words, and it needs an additional tool to recognize syllables. In Japanese, WikTra only works with Hiragana and Katakana scripts but not with Kanji (Chinese characters). We therefore combined WikTra with the Kuromoji[10] transliteration tool.

8 Geographic evidence

We exploit geographic information on languages in order to take into account the proximity of language speakers for the prediction of borrowing. Our hypothesis is that, even if in the last century lexical borrowing on a global scale has been faster than ever before, the effect of geographic distance is still a significant factor when applying cognate discovery to entire vocabularies. This effect is combined with orthographic similarity in Algorithm 1, in a way that geographic proximity increases the overall likelihood of word pairs being cognates, without being a necessary condition.

Our solution considers only the languages of the input words, computing a language proximity value between 0 and 1:

$$\text{GeoProx}(l_1, l_2) = \min(\frac{T_D}{\text{GeoDist}(l_1, l_2)}, 1.0) \quad (5)$$

The function $\text{GeoDist}(l_1, l_2)$ is an approximate 'geographic distance' between two languages l_1 and l_2, based on the geographical areas where the languages are spoken. The constant T_D corresponds to a *minimal distance*: if two languages are spoken within this distance then they have maximum geographic relatedness. T_D is empirically set as described in Sect. 9.3.

Distances between languages are provided by the WALS resource,[11] one of the most comprehensive language databases. WALS provides latitude and longitude coordinates for a language given as input. While a single coordinate returned for a language may in some cases be a crude approximation of linguistic coverage (e.g., Spanish is spoken both in Spain and in most countries of Latin America), even this level of precision was found to improve our evaluation results. For future versions of CogNet, we foresee the extension of our geographic evidence by a more fine-grained database on the geographic positions of language speakers.

9 Evaluation

This section describes how CogNet was evaluated on a diverse set of cognate corpora, and how its parameters were tuned to optimise results. With respect to the evaluation dataset used in Batsuren et al. (2019a), we have considerably extended

[10] https://github.com/atilika/kuromoji, accessed on 13/10/2018.

[11] https://wals.info.

the evaluation corpus size, and we have also incorporated a pre-existing cognate database into our evaluations. The creation of self-annotated evaluation datasets despite the existence of cognate databases was desirable due to the latter being either phonetic (and thus not usable for our purposes) or limited to very few language pairs (as the resource described below). Furthermore, performing validation simultaneously on very different datasets provides more precise and reliable results.

As a pre-existing cognate corpus, we used the *MFCogn*[12] (Most Frequent Cognates) dataset, available for four language pairs: English to Spanish (652 pairs), French (1281 pairs), Italian (1333 pairs), and Portuguese (1280 pairs). The total size of the dataset is thus 4546 word pairs.

Furthermore, in order to extend the coverage of our evaluation to all languages present in CogNet, we created a corpus[13] of 60 concepts with fully annotated sets of cognate groups. On average, a concept was represented in 103 languages by 139 words: 8353 words in total for the 60 concepts. The concepts were chosen from the *Swadesh basic word list* (Swadesh, 1952) and from the WordNet *core concepts* (Boyd-Graber et al., 2006). The words corresponding to these concepts were retrieved from the UKC. For each concept, we asked two language experts to find cognate clusters among its lexicalizations. The experts made their decisions based on online resources such as Wiktionary and the *Online Etymology Dictionary*.[14] Inter-annotator agreement (Cohen's kappa) was 95.27%. The resulting human-annotated dataset contained 8353 words, 62,752 pairs of cognate words and 587,357 pairs of non-cognate words. This set was significantly larger (by 80%) than the one we used in (Batsuren et al. 2019a). We divided this dataset into two equal parts: the first 30 concepts for hyperparameter tuning ("tuning") and the second 30 concepts for evaluation ("test").

9.1 Hyperparameter tuning

The goal was to optimise the algorithm with respect to three hyperparameters: the threshold of combined orthographic–geographic relatedness T_F (Sect. 4), the geographic proximity contribution parameter T_G, and the minimum distance T_D (Sect. 8).

We have created a three-dimensional grid with $T_F = [0.0; 1.0]$ (the higher the value, the more the strings need to be similar to be considered as cognates), $T_G = [0.0; 1.0]$ (the higher the value, the more geographic proximity is considered as evidence), and $T_D = [0.0; 22.0]$ (the unit of 1.0 corresponds to a distance of 1000 km, within which geographic relatedness is a constant maximum).

In this grid, we computed optimal values for each parameter (in increments of 0.01) based on performance on the hyperparameter tuning dataset described above. With these optimal settings, we evaluated all possible combinations of the various components of the cognate generation method, in order to understand their

[12] http://www.cognates.org/#resources.

[13] https://github.com/kbatsuren/CogNet/tree/master/evaluation.

[14] https://www.etymonline.com.

Table 1 Hyperparameter tuning and comparisons for the full system (top row), the full system with individual sources of evidence excluded to analyse their impact (rows 2–6), and baseline implementations (rows 7–10)

Methods	Parameters			Tuning			Test		
	T_F	T_G	T_D	R	P	F_1	R	P	F_1
Full system	0.71	0.06	1.5	84.13	96.29	89.80	82.88	95.62	**88.80**
Without DirectEvidence	0.67	0.03	1.5	54.49	97.83	69.99	50.54	94.59	65.87
Without TransitiveClosure	0.67	0.06	1.5	39.04	99.44	56.07	45.59	99.32	62.49
Without PhonSim	0.70	0.05	1.3	84.08	96.48	**89.85**	82.12	95.64	88.37
Without OrthSim	0.67	0.06	1.5	60.56	98.69	75.13	61.54	96.86	75.26
Without GeoProx	0.69	–	–	83.72	96.31	89.57	81.04	95.67	87.75
Baseln 1: LCS	0.58	–	–	14.47	95.75	25.14	19.76	96.40	32.80
Baseln 2: Consonant	–	–	–	11.27	97.21	20.20	16.05	94.19	27.42
Baseln 3: DirectEvidence	–	–	–	29.06	99.98	45.03	32.73	99.98	49.32

Values in bold are the overall highest F-measures obtained for tuning and test

relative contribution to the overall score. We favoured precision over recall, setting our minimum precision threshold to 95% and maximizing F-measure with respect to this constraint. The best settings (computed on the tuning dataset) as well as the corresponding precision—recall figures (computed on the test set) are reported in Table 1. The final precision results turned out to be even higher than our minimum threshold: 96.29% on the tuning and 95.62% on the test set.

Tuning results can be seen in Table 1. The optimal *geographic region* parameter T_D was computed to be 1.5: languages spoken within a distance of 1500 km share most of the cognates.

9.2 Evaluation results

In order to have a clear idea of the precision of CogNet and the performance of its components, we performed three separate evaluations:

- a systematic evaluation of precision and recall of the entire CogNet system and the impact of its resources on performance, using the self-annotated evaluation dataset;
- manual evaluation of precision by language experts of 400 randomly sampled cognate pairs not present in the self-annotated dataset;
- evaluation of precision and coverage against the external MFCogn dataset.

As it can be seen from the results below, all three evaluations provided very similar precision results of around 95%. The coherence of these values, which corresponds to our inter-annotator agreement, confirms the reliability of our evaluations.

Results over the Self-Annotated Corpus. We evaluated the effect of the various components of our method (direct etymological evidence, transitive closure,

Table 2 Evaluation results over the four language pairs of the MFCogn resource. Recall figures need to be interpreted in the context of UKC lexical incompleteness, indicated in the last column (in percentage of the number of synsets compared to the English Princeton WordNet

Language pair	Size	Precision (%)	Recall (%)	Voc. completeness (%)
English–French	1281	95.83	39.37	49.22
English–Portuguese	1280	94.91	39.95	35.46
English–Italian	1333	94.75	24.30	30.57
English–Spanish	682	93.09	37.63	32.36
Total	4546	94.86	34.86	37.67

phonetic similarity, orthographic similarity, geographic relatedness) on its overall performance. As baseline, we used two string similarity methods often applied to cognate identification (St Arnaud et al., 2017): *LCS*, i.e. the longest common subsequence ratio of two words (which we also use in Eq. 2), and *Consonant* (Turchin et al., 2010), which is a heuristic method that checks if the first three consonants of the words are identical. We also present as baseline our system having only gold standard etymological evidence as input. Results (Table 1) show that all components have contributed to the overall recall of 82.88% obtained at the precision level of 95.62%. The biggest impact came from transitive closure, then direct etymological evidence, then orthographic similarity, and then geographic proximity. Phonetic similarity has the lowest (yet still positive) impact on recall, which we attribute to the moderate amount of phonetic evidence collected so far. With this full setup we were able to generate 8,117,982 cognate pairs across 338 languages.

Results over the Random Sample. In order to cross-check the quality of the output, we randomly sampled 400 cognate pairs not covered by the self-annotated evaluation corpus and had them re-evaluated by the same expert annotators. Accuracy was found to be in the 93–97% range, in line with the goal of 95% we initially set in Sect. 9.3.

Results over the External Corpus. Precision and coverage results over MFCogn are shown in Table 2. Over all four language pairs, precision falls into the 93–96% range, the overall result being 94.86%. Again, this number is perfectly in line with the precision goal set by parameter tuning and with the results obtained by the two other evaluation methods. Recall is lower than what was obtained over the self-annotated corpus, which is explained by resource incompleteness: a significant proportion of words and meanings appearing in MFCogn are missing from the UKC vocabularies that provide the candidate word pairs to CogNet (see Table 2 for statistics on incompleteness). A special case of incomplentess concerns named entities: MFCogn contains name pairs such as *America–Amérique* while names are by design excluded from our source lexical database, as mentioned in Sect. 5. For this reason, recall figures obtained over MFCogn are not representative of the efficiency of the CogNet algorithm.

10 CogNet resources and their evolution

CogNet is designed to be a living and evolving resource, with subsequent releases extending its precision and coverage of languages and words, as well as following the evolution of the languages themselves (as far as this evolution is reflected in its input resources). We have so far produced three versions of CogNet, each one improving the previous version both by the extension of existing evidence and by the introduction of new kinds of evidence, as shown in the architecture design (Fig. 1, Sect. 3):

- CogNet v0, a preliminary version, used the UKC as its input lexical DB and relied only on direct etymological evidence from the Etymological WordNet and transitivity.
- CogNet v1, described in Batsuren et al. (2019a), still relied solely on the UKC but also included indirect evidence (transliteration, orthographic, and geographic).
- CogNet v2, presented in this paper, significantly increased its coverage thanks to extending the input lexical DB by about 800 thousand words retrieved from the PanLex resource. It also introduced a greatly extended etymological database as well as phonetic evidence, both extracted from Wiktionary.
- CogNet v3, finally, is under development, with further extensions of the input lexical database, new wordnets (Abiderexiti & Sun, 2019; Batsuren et al., 2019b; Loukachevitch & Gerasimova, 2019; Nair, 2019; Bella et al., 2020; Agostini et al., 2021; Bakay et al., 2021), as well as new sources of phonetic and geographic evidence (Lee et al., 2020), as mentioned in Sects. 7 and 8.

Table 3 provides a comparison among the versions in terms of precision and coverage (number of cognate word pairs found, number and percentage of words and concepts covered from the input lexical DB). The release of v1 resulted in a slight loss of precision, as expected due to the incorporation of indirect evidence. It however increased the number of cognates found by 50-fold (from 90k pairs to over half a million), and covering one-third of all words and three-fourth of all concepts in a cognate relationship. The release of v2 resulted in a further +152% extension of cognate coverage all the while improving precision by 1.68%.

All CogNet versions are distributed as simple tab-separated, freely downloadable text files, in the following tuple form:

$$(meaningID, w_1, l_1, w_2, l_2, evidence)$$

where *meaningID* is a meaning identifier representing the shared meaning of the cognate pair (which in versions 1–2 correspond to the Princeton WordNet English synset ID), w_1 and w_2 are the two words, l_1 and l_2 are their respective languages (expressed as ISO-639-3 codes). *Evidence* is an attribute describing the kind of linguistic evidence on which the prediction is based: *direct etymological*, *direct with transitivity*, or *indirect*. Cognates that were computed based on direct etymological

Table 3 CogNet versions and the respective input resources, types of evidence used, precision, number of cognate pairs, word coverage, and concept coverage

Ver	Lexical DB (# words)	Evidence	Prec.	Cogn	Words	Conc
v0	UKC (1.7M)	sem, ety	100.0%	0.06 M	0.09 M	9k
v1	UKC (1.7M)	sem, ety, orth, geo	93.94%	3.16 M	0.57 M	81k
v2	UKC+PanLex (2.5M)	sem+, ety+, orth, geo, phon	95.62%	8.41 M	1.08 M	91k
v3	UKC+PanLex (3M)	sem++, ety+, orth+, geo+, phon+	Work in progress			

evidence alone (29.3% of CogNet v2) correspond to the strict interpretation of cognacy used in historical linguistics, and is applicable in that field of research, under the condition that the underlying evidence (wordnets, the Etymological WordNet, and Wiktionary-based etymological knowledge) is accepted by the community. The set obtained after transitive closure of this strict subset is considerably larger (50.1% of CogNet v2) yet still compatible with the same strict interpretation due to the cognacy relationship being considered as symmetric and transitive (List, 2014). Lastly, cognates computed from indirect evidence also contain borrowings and other non-historically related word pairs. Due to their much larger coverage, they are more suitable for applications in computational linguistics.

11 Cognates and language diversity

While we expect CogNet to provide linguistic insights for both theoretical and applied research, we are just starting to exploit its richness. In this section we explore the contents of CogNet through the newly defined measures of *cognate distribution*, *cognate density*, and *cognate diversity*. We show how these measures provide insights into the geographic and genetic properties of cognates and, indirectly, of languages. We also show how we used CogNet to compute *lexical similarity* data and its visualization for over 27,000 language pairs. As we discuss below, beyond the insights they offer into the diversity of languages, these measures are also exploitable in cross-lingual NLP tasks.

11.1 Cognate distribution

We define the *cognate distribution* as a function

$$f^{C}_{\text{distr}}(d) = |\{(\langle w_1, l_1\rangle, \langle w_2, l_2\rangle)\}| \text{ such that } \text{GeoDist}(l_1, l_2) = d.$$

That is, for a given geographical distance d it returns the number of cognate pairs where the estimated distance of the speakers is d.

Figure 2 shows the plot of f^{C}_{distr}, alongside the plot of f_{distr} calculated over 'cognate candidates', i.e. all word pairs from the input corpus sharing a meaning (the word pairs that serve as input to the CogNet algorithm). Comparing the

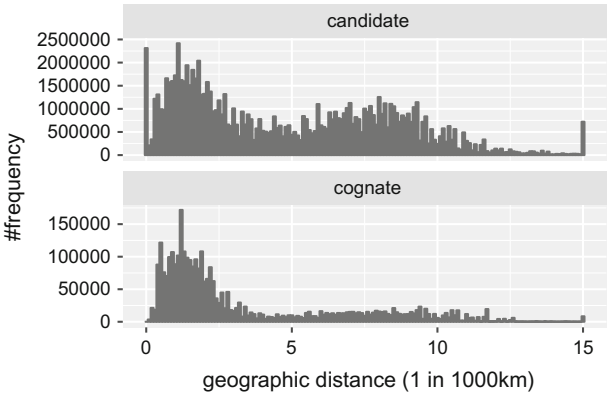

Fig. 2 Distribution of cognates in terms of the geographic distance of speakers (8.14 M pairs, at the bottom) and of all candidate word pairs (135 M, at the top)

geographic distribution of cognates to that of our entire input corpus, we observe that cognates demonstrate a much more pronounced locality: the vast majority of cognates is found within a distance of about 3000 km. Our interpretation of these results is that, by and large, the geographic proximity of speakers still has a major influence on modern lexicons, despite the globalising effects of the last centuries. Let us note that the GeoProx component of our algorithm is not causing this pronounced shift in the distribution, as it had a relatively minor overall contribution on the results (see the geographic factor $T_G = 0.06$ in Table 1). We computed Fig. 2 both with and without the GeoProx component, and the results did not show any significative difference.

11.2 Cognate density

Any per-language statistic computed over our cognate data is prone to be biased by the different levels of completeness of the lexicons used as input. As an abstraction from lexicon size, we introduce the notion of *cognate density*, f^C_{density}, of a language l, defined as the ratio of words of the language covered by at least one cognate pair:

$$f^C_{\text{density}}(l) = \frac{|\text{CognateWords}(l)|}{|\text{Words}(l)|}$$

In other terms, working with cognate densities allows us to characterise the 'cognate content' (as a percentage value) of each language independently of lexicon size. As a further attempt to avoid incompleteness bias, we limited our study of cognate densities to the 45 languages with a vocabulary size larger than 10,000 words.

Cognate densities for the 45 languages studied show a wide spread between languages with the highest density (the top five language being Indonesian: 60.80%, Czech: 59.05%, Catalan: 58.66%, Malay: 57.63%, and French: 57.25%) and those with the lowest (the bottom five languages being Thai: 7.87%, Arabic: 9.01%,

Table 4 Cognate density by language family, computed over the 45 largest-vocabulary languages

Family	Density (%)	Family	Density (%)
Malay	59.22	Greek	22.99
Romance	53.32	Niger-Congo	18.63
Slavic	36.67	Japanese	12.16
Indo-Aryan	36.08	Sino-Tibetan	11.22
Germanic	34.10	Mongolian	10.37
Basque	32.82	Persian	9.64
Dravidian	24.79	Arabic	9.01
Finno-Ugric	24.57	Thai	7.87

Persian: 9.64%, Mongolian: 10.37%, and Mandarin Chinese: 11.03%). The main factor behind high cognate density is the presence of closely related languages in our data: as Malay and Indonesian are mutually intelligible variants of essentially the same language, the existence of separate wordnets for the two naturally results in a high proportion of shared vocabulary. Inversely, languages on the other end of the spectrum tend not to have major living languages that are closely related. Let us note that transliteration mistakes may also be a reason for low cognate recall across languages with different scripts, especially in the case of scripts that are hard to transliterate, such as Chinese, Arabic, or Thai.

To verify these intuitions, we examined cognate densities for the 45 languages manually clustered into 16 language families (Table 4, the language name was kept for clusters of size 1). Families such as Malay, Romance, Slavic, or Indo-Aryan, well known for containing several mutually intelligible language pairs, came out on top, while families with fewer or mutually non-intelligible members at the bottom, such as Mongolian, Arabic, or Thai. The outlier is Basque that, despite being an isolate, is close to the resource-wide average cognate density of 33%.

11.3 Cognate diversity

We define the *cognate diversity* of a concept as

$$f^{C}_{\text{div}}(c) = 1 - \frac{|\text{LargestCognateSet}(c)|}{|\text{Languages}(c)|}.$$

where LargestCognateSet(c) is the largest set of words that express the concept c and that are cognates of each other. Note that each word in this set belongs to a different language. Languages(c), in turn, is the set of languages that lexicalise c. Cognate diversity captures how diverse or, on the contrary, universal is the lexicalisation of a given concept around the world. Words such as *coffee*, *tennis*, or *panda* were borrowed by almost all languages and thus these concepts have a cognate diversity close to zero, while the lexicalisations of *girl* or *tablecloth* are very heterogeneous around the world (see Table 5). Concepts that are lexicalized differently in each language (without a single cognate instance) would have the

Table 5 Cognate diversity of the eight most diverse (to the left) and least diverse or 'universal' (to the right) concepts

Concept	Diversity	Domain
Tadpole	0.970	Animal
Measles	0.955	Disease
Toadstool	0.953	Plant
Tablecloth	0.952	Artifact
Lass	0.951	Person
Light-year	0.945	Quantity
Girl	0.945	Person
Eyelash	0.944	Body part
Chlorine	0.061	Substance
Lithium	0.067	Substance
Neutron	0.077	Object
Gram	0.082	Quantity
Coffee	0.084	Food
Kilometer	0.089	Quantity
Guitar	0.097	Artifact
Gorilla	0.099	Animal

highest cognate diversity of 1; however, in our database not a single such concept could be found.

'Universal' concepts—and the cognate diversity measure that allows finding them—have practical uses in computational linguistics. As they are concepts with very similar lexicalisations in almost all languages, they can be exploited in machine translation and other cross-lingual applications. They can be used as seeds for unsupervised learning methods or for addressing the incompleteness of bilingual corpora, especially for under-resourced languages.

11.4 Lexical similarity

In comparative linguistics, the notion of lexical similarity refers to the relatedness of the lexicons of languages. The measure has been used, especially in the fields of lexicostatistics and glottochronology, to infer or verify the genetic relatedness between languages based on their vocabularies. The standard approach to formalise the similarity of two lexicons is to count the cognates shared between them over a controlled set of lexemes, sometimes also taking into account the phonetic similarity of words (Wichmann et al. 2010).

The large-scale data provided by CogNet can also be used to compute lexical similarities. The difference with respect to data used in lexicostatistics, however, is that CogNet data provides a contemporary synchronic rather than a diachronic perspective on language similarity: it is based on modern orthography (as opposed to phonetics that take sound changes into account) and it covers both historical and recent vocabulary. Consequently, results obtained from CogNet provide evidence on the similarity of lexicons *as they are today* as opposed to historical relatedness.

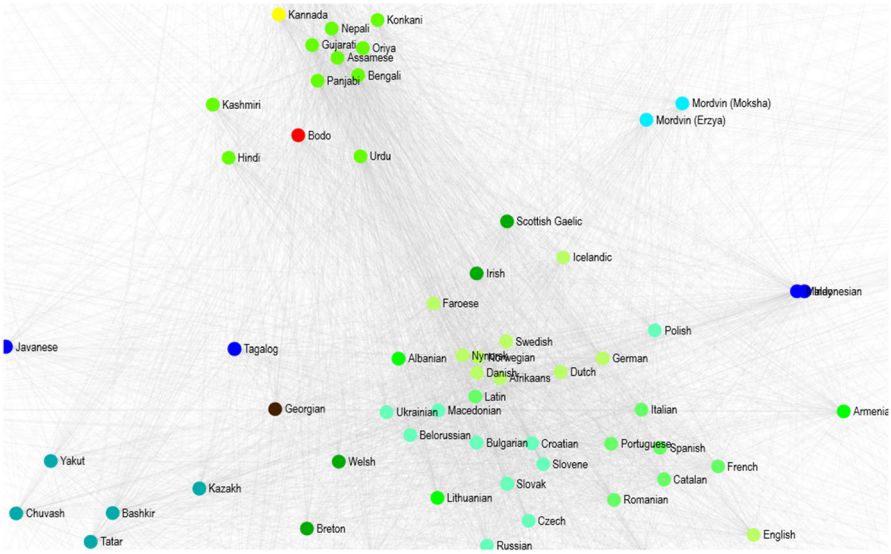

Fig. 3 A detail from a lexical similarity graph computed from CogNet data

Thus, beyond the study of the lexicons themselves, CogNet-based similarities are more adapted to cross-lingual computational NLP applications, such as to estimate the performance of cross-lingual transfer across word embeddings, than to studies in historical linguistics.

We have computed a lexical similarity database using CogNet data, providing similarity scores for over 27,000 language pairs. The computation was based on counting the number of cognates in each language pair, normalized by lexicon sizes. A classic graph-based visualization of the similarity data obtained can be seen in Fig. 3, where nodes represent languages, colors the language families to which they belong, and the distance between two nodes is proportional to their lexical similarity. Both the similarity database and the corresponding dynamic graph visualization are freely accessible online (Bella et al., 2021).[15]

12 Use case: bilingual lexicon induction

As an example of use of CogNet for research in computational linguistics, we present experiments on *bilingual lexicon induction* (BLI) (Haghighi et al., 2008). BLI is frequently used as a standard evaluation method for cross-lingual word embeddings produced through transfer learning, which has recently become a highly-researched topic. The goal of BLI is to learn word translations from word vector representations of bilingual non-parallel corpora. State-of-the-art supervised and semi-supervised methods use a small-to-medium-size bilingual dictionary as

[15] http://ukc.datascientia.eu/lexdist.

Table 6 Precision@1 for BLI on the VecMap tool with a mutli-step framework of linear transformations

Training dictionary	EN–IT (%)	EN–DE (%)	EN–FI (%)	EN–ES(%)
VecMap	48.40	49.20	36.10	**39.20**
Wordnet	42.53	44.20	29.28	33.93
CogNet	47.20	46.73	29.99	36.20
VecMap + Wordnet	46.53	48.20	32.94	37.80
VecMap + CogNet	**48.93**	**51.13**	**37.29**	38.81

Bold values indicates the highest precision scores obtained for the given language pair

Table 7 Dictionary size and average number of translations per source word in the training dictionaries

Training dictionary	EN–IT		EN–DE		EN–FI		EN–ES	
	Size	#Trans	Size	#Trans	Size	#Trans	Size	#Trans
VecMap	5k	1.24	5k	1.29	5k	1.39	5k	1.27
CogNet	16k	1.02	15k	1.03	19k	1.04	25k	1.02
VecMap + CogNet	21k	1.16	20k	1.15	24k	1.16	26k	1.15
Wordnet	67k	2.72	66k	2.14	284k	2.84	71k	2.29
VecMap + Wordnet	72k	2.68	71k	2.15	289k	2.86	76k	2.32

seed to compute cross-lingual mappings between the embedded vector spaces of the corpora of the two languages (Artetxe et al., 2016, 2018b).

Our goal was to evaluate the performance of CogNet used as seed dictionary, with respect to standard alternatives: sense-aligned wordnets in the five languages, as well as the well-known evaluation dataset introduced by Dinu and Baroni (2015) and known in the community as the *VecMap dataset*. We have used the *VecMap* tool[16] (Artetxe et al., 2018a) to train and evaluate cross-lingual mappings over four language pairs: from English to Italian, German, Finnish, and Spanish. Table 7 provides information on the training datasets used. The intuition behind using cognates as seed is that they typically represent high-quality translations of primary word meanings, which are by definition the focus of the BLI task.

Our evaluation results are presented in Table 6. CogNet used as a seed dictionary on its own consistently outperformed the much larger wordnets for all language pairs, except for the VecMap dataset. However, when combined with VecMap, it achieved the highest precision for three language pairs out of four, again before wordnets.

It is interesting to observe that the smallest resources (VecMap and CogNet) performed best, while the much larger and more complete wordnets achieved consistently worse results (see Table 7 for dictionary sizes). Our hypothesis is that the high degree of polysemy and synonymy present in wordnets has a detrimental effect on the BLI results: wordnets often generate more than one, and sometimes

[16] https://github.com/artetxem/vecmap.

even a very high number of possible translations for each English word, most of which, while formally correct, will be considered as false positives by the BLI evaluation that concentrates on the most common word senses and translations. For example, the word *cat* has no less than ten meanings in the English wordnet, and its most common meaning has three synonymous translations into Italian. To underpin this hypothesis, we have computed the average number of translations per English word in all resources (see Table 7). All wordnets have an average number of translations greater than 2, while in VecMap and CogNet this value is close to 1.

We consider these first experimental results as evidence of the usefulness of high-quality multilingual resources, and of CogNet in particular, in research and applications in cross-lingual tasks.

13 Conclusions

We have presented a general method for building a large-scale and free cognate database using existing cross-lingual lexicons, as well as a combination of orthographic, phonetic, semantic, etymological, and geographic evidence. After considerable extension, the current second release of CogNet contains 152% more (8.1 M) sense-tagged word pairs in 338 languages and 35 scripts, at a precision of 95%. Future releases with even greater coverage are already in preparation. Finally, as an illustration of how high-quality cross-lingual resources such as CogNet have a high potential of reuse in corpus-driven cross-lingual tasks in computational linguistics and NLP, we managed to obtain improvements over state-of-the-art results on a use case of bilingual lexicon induction via cross-lingual mapping of word embeddings.

Acknowledgements The first author is supported by the Early Stage Researcher project, funded by the National University of Mongolia under Grant P2019-3716.

Funding Open access funding provided by Università degli Studi di Trento within the CRUI-CARE Agreement.

References

Abiderexiti, K., & Sun, M. (2019). Construction of an English-Uyghur wordnet dataset. In *China national conference on Chinese computational linguistics* (pp. 382–393). Springer, Berlin

Agostini, A., Usmanov, T., Khamdamov, U., Abdurakhmonova, N., & Mamasaidov, M. (2021). Uzwordnet: A lexical-semantic database for the uzbek language. In *Proceedings of the 11th Global Wordnet conference* (pp. 8–19)

Artetxe, M., Labaka, G., & Agirre, E. (2016). Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 2289–2294).

Artetxe, M., Labaka, G., Agirre, E. (2018a). Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Thirty-second AAAI conference on artificial intelligence*.

Artetxe, M., Labaka, G., & Agirre, E. (2018b). A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th annual meeting of the association for computational linguistics* (Vol. 1: Long Papers, pp. 789–798).

Bakay, Ö., Ergelen, Ö., Sarmış, E., Yıldırım, S., Arıcan, B. N., Kocabalcıoğlu, A., Özçelik, M., Sanıyar, E., Kuyrukçu, O., & Avar, B., et al. (2021). Turkish wordnet kenet. In *Proceedings of the 11th global wordnet conference* (pp. 166–174).

Batsuren, K., Bella, G., & Giunchiglia, F. (2019a). Cognet: A large-scale cognate database. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 3136–3145).

Batsuren, K., Ganbold, A., Chagnaa, A., & Giunchiglia, F. (2019b). Building the mongolian wordnet. In *Proceedings of the 10th global WordNet conference* (pp. 238–244).

Beinborn, L., Zesch, T., & Gurevych, I. (2013). Cognate production using character-based machine translation. In *Proceedings of the sixth international joint conference on natural language processing* (pp. 883–891).

Bella, G., Zamboni, A., & Giunchiglia, F. (2016). Domain-based sense disambiguation in multilingual structured data. In *The diversity workshop at the 22nd European conference on artificial intelligence* (*ECAI* 2016).

Bella, G., Giunchiglia, F., & McNeill, F. (2017). *Language and Domain Aware Lightweight Ontology Matching*. Web Semantics: Science, Services and Agents on the World Wide Web.

Bella, G., McNeill, F., Gorman, R., Ó Donnaíle, C., MacDonald, K., Chandrashekar, Y., Alhakim Freihat, A., & Giunchiglia, F. (2020). A major wordnet for a minority language: Scottish gaelic. In *Proceedings of the 12th international conference on language resources and evaluation* (*LREC* 2020).

Bella, G., Batsuren, K., & Giunchiglia, F. (2021). A database and visualization of the similarity of contemporary lexicons. *Proceedings of the 24th international conference on text, speech and dialogue*. Springer.

Bond, F., & Foster, R. (2013). Linking and extending an open multilingual wordnet. In *ACL* (Vol. 1, pp. 1352–1362).

Bouckaert, R., Lemey, P., Dunn, M., Greenhill, S. J., Alekseyenko, A. V., Drummond, A. J., et al. (2012). Mapping the origins and expansion of the Indo-European language family. *Science, 337*(6097), 957–960.

Boyd-Graber, J., Fellbaum, C., Osherson, D., & Schapire, R. (2006). Adding dense, weighted connections to wordnet. In *Proceedings of the third international WordNet conference, Citeseer* (pp. 29–36).

Chakravarthi, B. R., Rajasekaran, N., Arcan, M., McGuinness, K., O'Connor, N. E., & McCrae, J. P. (2020). Bilingual lexicon induction across orthographically-distinct under-resourced dravidian languages. In *Proceedings of COLING 2020*.

Ciobanu, A. M., & Dinu, L. P. (2020). Automatic identification and production of related words for historical linguistics. *Computational Linguistics*, 667–704.

Crowley, T., & Bowern, C. (2010). *An introduction to historical linguistics*. Oxford: Oxford University Press.

De Melo, G. (2014). Etymological wordnet: Tracing the history of words. In *LREC, Citeseer* (pp. 1148–1154).

Dinu, G., & Baroni, M. (2015). Improving zero-shot learning by mitigating the hubness problem. In: *Workshop track of international conference on learning representations* (Vol. abs/1412.6568).

Giunchiglia, F., Batsuren, K., & Bella, G. (2017). Understanding and exploiting language diversity. In *Proceedings of the twenty-sixth international joint conference on artificial intelligence* (*IJCAI*-17) (pp. 4009–4017).

Giunchiglia, F., Batsuren, K., & Freihat, A. A. (2018). One world—seven thousand languages. In *Proceedings 19th international conference on computational linguistics and intelligent text processing, CiCling2018*, 18–24 March 2018.

Greenhill, S. J., Blust, R., & Gray, R. D. (2008). The austronesian basic vocabulary database: from bioinformatics to lexomics. *Evolutionary Bioinformatics, 4*, EBO–S893.

Haghighi, A., Liang, P., Berg-Kirkpatrick, T., & Klein, D. (2008). Learning bilingual lexicons from monolingual corpora. In *Proceedings of ACL-08: Hlt* (pp. 771–779).

Hauer, B., & Kondrak, G. (2011). Clustering semantically equivalent words into cognate sets in multilingual lists. In *Proceedings of 5th international joint conference on natural language processing* (pp. 865–873).

Hermjakob, U., May, J., & Knight, K. (2018). Out-of-the-box universal romanization tool uroman. In *Proceedings of ACL 2018, system demonstrations* (pp. 13-–8).

Jäger, G. (2013). Phylogenetic inference from word lists using weighted alignment with empirically determined weights. *Language Dynamics and Change, 3*(2), 245–291.

Jäger, G. (2018). Global-scale phylogenetic linguistic inference from lexical resources. *Scientific Data, 5,* 180189. https://doi.org/10.1038/sdata.2018.189.

Jäger, G., List, J. M., & Sofroniev, P. (2017). Using support vector machines and state-of-the-art algorithms for phonetic alignment to identify cognates in multi-lingual wordlists. In *Proceedings of the 15th conference of the European chapter of the Association for Computational Linguistics* (Vol. 1, Long Papers, pp. 1205–1216).

Kamholz, D., Pool, J., & Colowick, S. M. (2014). Panlex: Building a resource for panlingual lexical translation. In *LREC* (pp. 3145–3150).

Kanojia, D., Dabre, R., Dewangan, S., Bhattacharyya, P., Haffari, G., & Kulkarni, M. (2020). Harnessing cross-lingual features to improve cognate detection for low-resource languages. In *Proceedings of the 28th international conference on computational linguistics* (pp. 1384–1395).

Kanojia, D., Sharma, P., Ghodekar, S., Bhattacharyya, P., Haffari, G., & Kulkarni, M. (2021). Cognition-aware cognate detection. In *Proceedings of the 16th conference of the European chapter of the Association for Computational Linguistics: Main Volume* (pp. 3281–3292).

Karimi, S., Scholer, F., & Turpin, A. (2011). Machine transliteration survey. *ACM Computing Surveys (CSUR), 43*(3), 17.

Kondrak, G. (2000). A new algorithm for the alignment of phonetic sequences. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference* (pp. 288–295). Association for Computational Linguistics.

Kondrak, G., Marcu, D., & Knight, K. (2003). Cognates can improve statistical translation models. In *Proceedings of the 2003 conference of the North American chapter of the Association for Computational Linguistics on human language technology: Companion volume of the proceedings of HLT-NAACL 2003–short papers-Volume 2* (pp. 46–48). Association for Computational Linguistics.

Lee, J. L., Ashby, L. F., Garza, M. E., Lee-Sikka, Y., Miller, S., Wong, A., McCarthy, A. D., & Gorman, K. (2020). Massively multilingual pronunciation modeling with wikipron. In *Proceedings of the 12th language resources and evaluation conference* (pp. 4223–4228).

List, J. M. (2012). Lexstat: Automatic detection of cognates in multilingual wordlists. In *Proceedings of the EACL 2012 joint workshop of LINGVIS & UNCLH* (pp 117–125). Association for Computational Linguistics.

List, J. M. (2014). Sequence comparison in historical linguistics. PhD thesis, Düsseldorf University Press.

List, J. M. (2019). Automated methods for the investigation of language contact, with a focus on lexical borrowing. *Language and Linguistics Compass, 13*(10), e12355.

List, J. M. (2019). Automatic inference of sound correspondence patterns across multiple languages. *Computational Linguistics, 45*(1), 137–161.

Loukachevitch, N., & Gerasimova, A. (2019). Linking Russian wordnet ruwordnet to wordnet. In *Proceedings of the 10th global wordnet conference* (*GWC* 2019) (pp. 64–71).

Miller, G. A. (1995). Wordnet: A lexical database for English. *Communications of the ACM, 38*(11), 39–41.

Nair, N. C., Velayuthan, R. S., & Batsuren, K. (2019). Aligning the indoWordNet with the Princeton WordNet. In *Proceedings of the 3rd international conference on natural language and speech processing* (pp. 9–16).

Navigli, R., & Ponzetto, S. P. (2010). Babelnet: Building a very large multilingual semantic network. In *Proceedings of the 48th annual meeting of the association for computational linguistics* (pp. 216–225). Association for Computational Linguistics.

Rama, T., Wahle, J., Sofroniev, P., & Jäger, G. (2017). Fast and unsupervised methods for multilingual cognate clustering. arXiv:170204938.

Rama, T., List, J. M., Wahle, J., & Jäger, G. (2018). Are automatic methods for cognate detection good enough for phylogenetic reconstruction in historical linguistics? In *Proceedings of the 2018 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies* (Vol. 2, pp. 393–400).

Sennrich, R., Haddow, B., & Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th annual meeting of the Association for Computational Linguistics* (Vol. 1, pp. 1715–1725).

St Arnaud, A., Beck, D., & Kondrak, G. (2017). Identifying cognate sets across dictionaries of related languages. *Proceedings of the EMNLP, 2017,* 2519–2528.

Swadesh, M. (1952). Lexico-statistic dating of prehistoric ethnic contacts: With special reference to North American Indians and Eskimos. *Proceedings of the American Philosophical Society, 96*(4), 452–463.

Tsvetkov, Y., & Dyer, C. (2015). Lexicon stratification for translating out-of-vocabulary words. In Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing (Vol. 2: Short Papers, pp. 125–131).

Turchin, P., Peiros, I., & Gell-Mann, M. (2010). Analyzing genetic connections between languages by matching consonant classes. *Journal of Language Relationship, 5,* 117–126.

Wichmann, S., Müller, A., Velupillai, V., Brown, C. H., Holman, E. W., Brown, P., Sauppe, S., Belyaev, O., Urban, M., & Molochieva, Z., et al. (2010). The ASJP database (version 13). http://www.emailevampgde/wichmann/ASJPHomePagehtm3.

Wu, W., & Yarowsky, D. (2018). Creating large-scale multilingual cognate tables. In *Proceedings of the eleventh international conference on language resources and evaluation* (*LREC*-2018).