Language Diversity: Visible to Humans, Exploitable by Machines

Gábor Bella
University of Trento
via Sommarive, 5, 38123 Trento, Italy
`gabor.bella@unitn.it`

Erdenebileg Byambadorj
University of Trento
via Sommarive, 5, 38123 Trento, Italy
`e.byambadorj@unitn.it`

Yamini Chandrashekar
University of Trento
via Sommarive, 5, 38123 Trento, Italy
`yamini.chandrashekar@unitn.it`

Khuyagbaatar Batsuren
National University of Mongolia
NUM 1, 14200, Ulaanbaatar, Mongolia
`khuyagbaatar@num.edu.mn`

Danish Asghar Cheema
University of Trento
via Sommarive, 5, 38123 Trento, Italy
`danish.cheema@unitn.it`

Fausto Giunchiglia
University of Trento
via Sommarive, 5, 38123 Trento, Italy
`fausto.giunchiglia@unitn.it`

Abstract

The *Universal Knowledge Core* (UKC) is a large multilingual lexical database with a focus on language diversity and covering over two thousand languages. The aim of the database, as well as its tools and data catalogue, is to make the abstract notion of linguistic diversity visually understandable for humans and formally exploitable by machines. The UKC website lets users explore millions of individual words and their meanings, but also phenomena of cross-lingual convergence and divergence, such as shared interlingual meanings, lexicon similarities, cognate clusters, or lexical gaps. The *UKC LiveLanguage Catalogue*, in turn, provides access to the underlying lexical data in a computer-processable form, ready to be reused in cross-lingual applications.

1 Introduction

A recent challenge in computational linguistics has been the development of efficient multilingual and cross-lingual techniques for language understanding and processing. In terms of solutions, a mainstream, yet often implicit assumption has been that shared meaning unites languages beyond superficial differences in lexicon and grammar: after all, humankind on the whole has been successful in getting ideas across linguistic borders. Hence the recent trend of massively multilingual resources—lexical databases, cross-lingual transfer matrices, pre-trained multilingual language models—exploiting a common meaning-based mapping across linguistic units.

Linguistic diversity remains, nevertheless, a key concept insomuch as it refers to deep-running, irreducible, meaning-level differences across languages and underlying cultural concepts. To take real examples from state-of-the-art machine translation, syntactically correct but semantically absurd outputs such as *'my older brother is younger than me'* or *'this raw rice is tasty'* are not rare exceptions but recurrent consequences of diversity: the diverging ways languages express culturally significant concepts such as *brother* or *rice*. While phenomena such as lexical gaps (Lehrer, 1970), culturally diverse terminology, or the varying relevance of the notion of *word* itself across languages are not unfamiliar to us computational linguists, this intuitive and high-level understanding is hard to translate into actual 'diversity-aware' computational applications, not the least because of the lack of formal datasets that would provide such information.

Lexical typology has described and catalogued many of such phenomena (Koptjevskaja-Tamm et al., 2015). A few online databases also provide contrastive data and visualisations, sometimes over thousands of languages (Dryer and Haspelmath, 2013; Rzymski et al., 2020; Holman et al., 2011; Arora et al., 2021). The These databases are rarely used in the NLP community, probably because they are often targeted towards historical linguistics and use phonetic representations of words or are limited to a few hundred core concepts. Yet, our position is that typological data can and should be reused for computational purposes, provided that they are meaningfully integrated with existing resources on
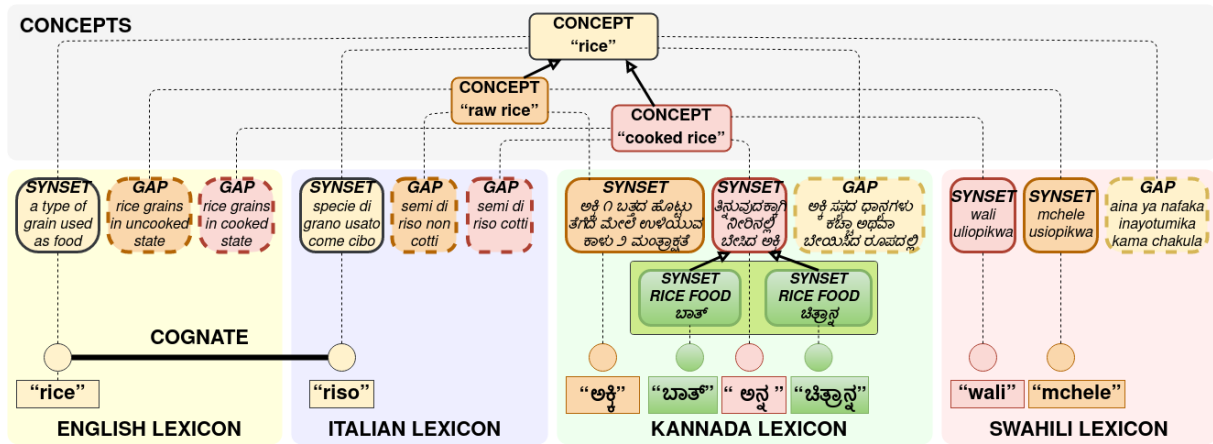
Figure 1: Structural elements in the UKC lexical database for representing cross-lingual unity and diversity.

contemporary language.

Computationally-oriented resources that address language diversity or linguistic typology have so far been concentrating on multilingual morphosyntax (Ponti et al., 2019; Batsuren et al., 2021b; Nivre et al., 2016). On diversity in lexical semantics, only a few studies (Giunchiglia et al., 2017) and sporadic data have been available for specific languages, such as a few hundred lexical gaps in Hebrew (Ordan and Wintner, 2007) or in Italian (Pianta et al., 2002). Large-scale multilingual lexical databases (MLDB), such as BabelNet (Navigli and Ponzetto, 2012) or the Open Multilingual Wordnet (Bond and Foster, 2013), have so far ignored phenomena related to language diversity and have concentrated on shared meaning.

The *Universal Knowledge Core* (UKC) database and system presented in this paper provides computer-readable cross-lingual lexical data, covering both the common and the diverse among more than a thousand lexicons. The data is being made available for download from the *UKC LiveLanguage* catalogue[1], while the *UKC website*[2] provides a set of interactive tools that allow both high-level visualisations and an in-depth exploration of diversity data. The rest of the paper provides an overview of the UKC database structure and contents, the online tools, and the data catalogue.[3]

2 A Multilingual Lexical Database on Language Diversity

Among existing large-scale MLDBs, those with a published formal, computer-exploitable data

model—such as the Open Multilingual WordNet, BabelNet, or EuroWordNet (Vossen, 1997)—concentrate solely on *language unity*, i.e. shared supra-lingual meaning, through linking together words with the same meaning across languages. The UKC offers a richer, two-layered representation of language unity, as well as introducing *language diversity* as formal data, both in terms of lexical model and actual content.

The UKC data model, the theoretical underpinnings of which have been exposed in (Giunchiglia et al., 2018), is illustrated in Figure 1. On the top of the figure, a supra-lingual *concept layer* contains hierarchies of concepts that represent lexical meaning shared across languages. Concepts thus act as bridges across languages. The only criterion for a concept to be present in the concept layer is that it is lexicalised by at least one language.

The bottom *lexicon layer* consists of language-specific lexicons. As in other lexical databases, these provide lexicalisations for concepts, such as the English *'rice'* and the Italian *'riso'* for the concept of *rice* in Figure 1. Beyond lexicalisations, however, the UKC lexicons also provide rich cross-lingual information on language unity and diversity through the additional constructs detailed below.

Lexical gaps. As evoked in the introduction, when unrecognised, lexical untranslatability can decrease the performance of cross-lingual applications. Syntactically correct yet meaningless Google translations, such as the Hungarian sentence

'A bátyám három évvel fiatalabb nálam,'

meaning 'my older brother is three years younger than me,' are systematically produced due to non-existent equivalent translations, in this case for the
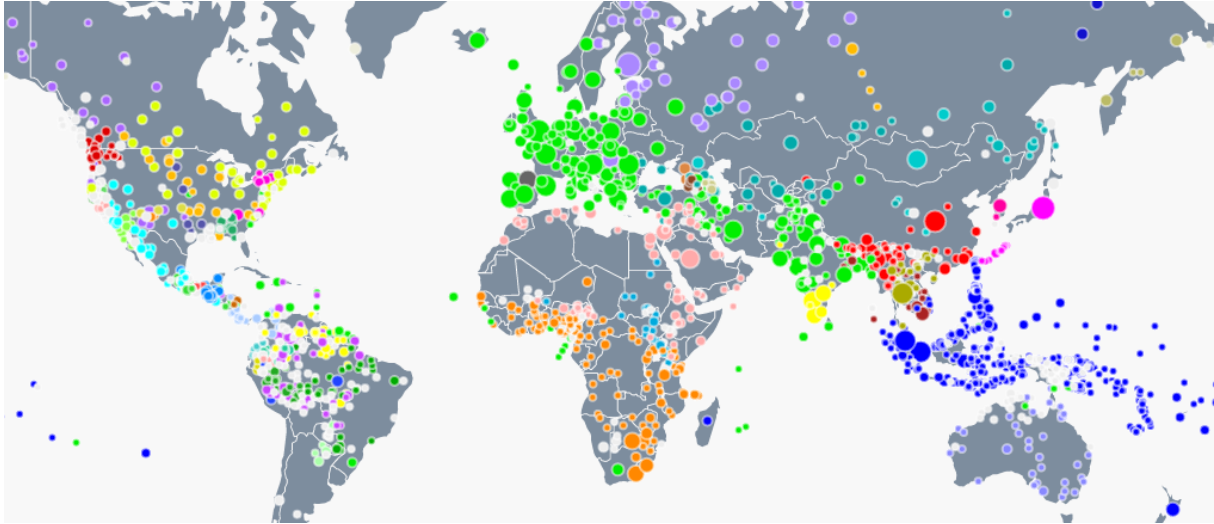
Figure 2: Lexicons in the UKC: circle sizes indicate lexicon size and their colour the language family (phylum).

word *brother* in Hungarian.[4] Likewise, as shown in Figure 1, English has no single word for *raw, uncooked rice* while Swahili—and many other languages, cf. (Joo, 2021)—has no equivalent for the general term *rice*. The UKC provides evidence of untranslatability by representing lexical gaps inside lexicons. Such information can be used, among others, to indicate the absence of equivalent terms to downstream cross-lingual applications.

Cross-lingual sense relations. Beyond providing shared word meanings as other MLDBs do, the UKC represents a richer set of interlingual connections between word senses. For example, in Figure 1, the English *'rice'* and the Italian *'riso'* are connected through a *cognate* relationship. Cognates are words in different languages that sound the same and have the same (or similar) meaning due to a common etymological origin. Cognates are key indicators of language unity on the lexical level, i.e. of the cross-lingual similarity of lexicons. Such information can thus be exploited e.g. as seeds in cross-lingual tasks such as bilingual lexicon induction (Batsuren et al., 2021a).

Metadata on language diversity. Beyond standard typological metadata such as language phylogeny or the geographical locations of speakers, the UKC also integrates cross-linguistic metadata computed from its own lexico-semantic content. Based on cross-lingual cognate relationships, we computed large-scale *lexicon similarity* data across 27 thousand language pairs over 331 languages. Lexicon similarity (Bella et al., 2021) formally

characterises the extent to which the vocabularies of two languages 'resemble each other', taking differing writing systems and orthographies into account. This metric has, in our view, a better potential in predicting the success of cross-lingual tasks (such as transfer learning or joint supervised training) than language phylogeny, as it is based on the overlaps of contemporary lexicons as opposed to historical relatedness.

Language-specific word meanings. Diversity also means acknowledging our partial understanding of how specific languages conceptualise lexical meaning and the ultimate impossibility of an exhaustive interlingual model. The UKC is the only lexical database to allow the co-existence of shared and language-specific word meaning hierarchies, inside the concept layer and the lexicons, respectively. Figure 1 shows culture-specific words and meanings (of rice-based foods) represented inside the Kannada lexicon, not yet integrated into the shared concept layer.

Language-specific lexical relations. Lexical relations within individual languages are sometimes part of lexical databases, such as antonymy or derivation in the Princeton WordNet (Miller, 1998). The UKC integrates derivational relationships for 20 languages, introduces relation types not typically part of lexical databases—such as *metonym-of* or *homograph-of*—and provides corresponding relation instances.

Table 1 shows the current contents of the UKC (as of January 2022) in terms of the data types

---

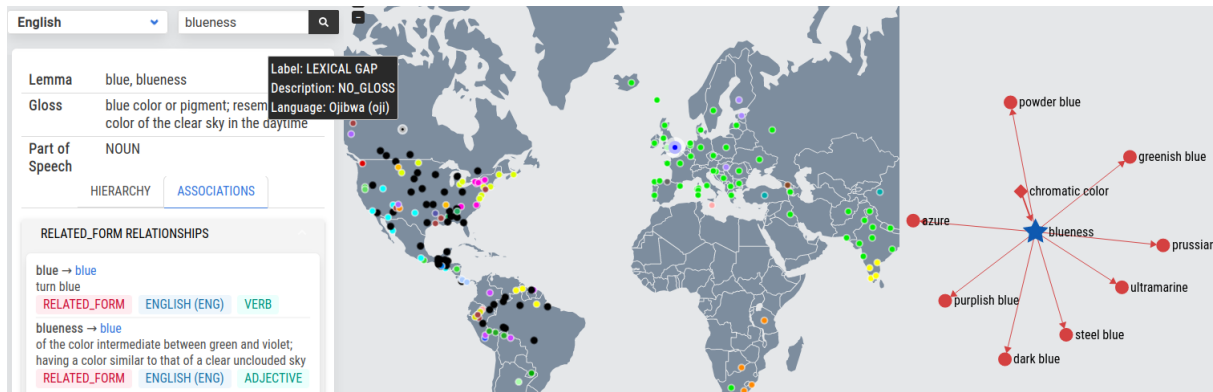[4]Apart from the laborious and thus rarely used *fiútestvér*.

Figure 3: Exploring the concept of *blueness* as lexicalised in the English language (left), in the world (middle), and as part of the supra-lingual concept hierarchy (right).

enumerated above. Concepts and concept relations were initially derived from the Princeton WordNet, as in all other MLDBS, but then extended with 400 new concepts and 490 new relations. We obtained lexicalisations from Wiktionary, NorthEuraLex (Dellert et al., 2020), the *Native Languages of the Americas*[5], as well as from 28 monolingual wordnets cited at the end of our paper. Lexical gaps were mostly obtained from original research (about 33k from diversity-rich domains such as kinship and colours, but also 600 gaps from (Bella et al., 2020)), and over a thousand gaps from the few third-party resources providing such information (Pianta et al., 2002; Ordan and Wintner, 2007). We computed cross-lingual sense relations from UKC data, reusing our method published in (Batsuren et al., 2019a, 2021a). Language-specific sense relations were obtained in a minor part from wordnets providing such data (48k relations), and in a major part from our own research on multilingual morphology (Batsuren et al., 2021b) (770k derivations in 16 languages) and metonymy (25k metonyms in 191 languages). Finally, lexicon-level metadata on language diversity combines online sources (Dryer and Haspelmath, 2013) with results of our own research on the similarity of lexicons (Bella et al., 2021).

3 Exploring Diversity Data

The website of the UKC database provides browseable online access to the full database contents, data visualisation tools, extensive information on related projects, publications, source materials, as well as example downstream services, such

---

[5] http://www.native-languages.org

Content type	Data size
Languages	2,176
Concepts	106k
Concept relations	109k
Lexicalisations (word senses)	2.8M
Lexical gaps	35k
Cross-lingual sense relations	8M
Language-specific relations	840k
Lexicon-level diversity metadata	30k

Table 1: UKC contents and data sizes.

as word translation between any two languages or multilingual word sense disambiguation (to be released soon).

A major feature of the website is the interactive exploration of lexicons and diversity data. The user can browse: (1) linguistic metadata of the 2k lexicons, selecting the language from an interactive map (Figure 2) or by name; (2) within a language, all meanings of a word typed in by the user; and (3) lexicalisations and gaps of a concept in the current language and in all languages of the world.

A screenshot of the last—and richest—concept exploration functionality, taking the example concept of *blueness*, is provided in Figure 3. On the left-hand side of the screen, details are provided on the lexicalisation of the concept in the current language, such as synonyms, definition, part of speech, as well as lexical relationships to other word senses (e.g. derivations, metonyms, cognates in other languages). The middle part of the screen shows an interactive clickable map of all languages that either lexicalise the concept in the database or, on the contrary, *are known not to lexicalise it*. The colour-coded dots (indicating the language family while 'black holes' stand for gaps) thus provide an

instant global typological overview for the concept selected, e.g. from Figure 3 one can see that a lot of languages in the Americas do not lexicalise blue as a separate colour. Some language do not appear on the map due to lexicon incompleteness: for those languages the UKC has no information whether they lexicalise *blueness* or not.

The right-hand side, finally, shows the concept in the context of the lexico-semantic concept hierarchy, displayed as a graph. A detail of the full graph is shown, including the currently observed concept *blueness*, the parent (broader), and the child (narrower) concepts. Other lexico-semantic relationships (e.g. meronymy and metonymically related concepts) are also shown when they exist. While for usability reasons the graph only displays a part of the full hierarchy, it is interactive, allowing the entire concept graph to be explored in the currently selected language. Changing the language is as simple as clicking on the map or selecting it from the drop-down in the upper left corner of the screen. Colours in the graph are indicative of language diversity: they show whether a concept is lexicalised in the current language (dark-coloured nodes), are missing from its lexicon (light-coloured nodes), or are lexical gaps (black nodes).

4 Visualising Language Diversity

Beyond the fine-grained word and concept exploration presented in the previous section, the UKC website also offers visualisation tools that allow humans to grasp diversity both in its globality and from different angles. Currently the following tools are provided, three of which we present below: (1) cognate diversity clusters; (2) colexifications; (3) a gap explorer for a fixed set of domains that are lexically diverse; (4) lexical similarity graphs; and (5) visual statistics.

Cognate diversity clusters. This tool shows *cognate clusters* on the map for a given concept selected by the user, computed from cognate data inside the UKC. In Figure 4, the concept of *fish* is selected: each dot represents a lexicon that contains a word for *fish*. Two dots are of the same colour if the two words are cognates of each other. For example, the English *'fish'* and the Italian *'pesce'* are within one cognate cluster (in light green in the figure) while the Hungarian *'hal'* and the Finnish *'kala'* are in another cluster (in turquoise). The number and distribution of clusters for a given concept provide information about its universality or

diversity: *coffee* is a so-called *universal concept* while *woman* is an extremely diverse one.

Lexical gap explorer. Certain domains—such as kinship, food, colours, or body parts—are known by linguists to be lexically diverse, for reasons related to culture, geography, but also grammar and other factors (Lehrer, 1970). The *gap explorer tool* displays a full concept hierarchy for a domain or subdomain selected by the user. Figure 5 shows the UKC concept structure of the subdomain of *siblings* from the *kinship* domain. For the language selected (Danish in the figure), the tool displays existing lexicalisations, indicates incompleteness (missing word), and provides known lexical gaps. This allows for quick comparisons of how different languages lexicalise (or not) a given domain. For example, for the *cousins* subdomain that consists of 67 concepts, English only lexicalises the root concept *cousin* with all other concepts as gaps, while South Indian languages provide no less than 16 distinct words depending on the age, sex, and lineage (patrilineal/matrilineal) of the cousin.

Lexical similarity graphs. Relying on the extensive lexical data inside the UKC, we compute pairwise similarities between languages based on the amount of shared cognates, using the method described in our recent paper (Bella et al., 2021). In order to interpret the resulting similarity data for humans, i.e. provide a global overview of lexical similarity, we compute a dynamic graph visualisation where nodes are languages and edge lengths are proportional to lexical similarities. The graph computation relies on a physical model of attraction and repulsion among nodes, using the *ForceAtlas2* library (Jacomy et al., 2014). We provide two distinct colourings for the same graph: one based on language families (shown in Figure 6) and the other based on geographical distance. These graphs visualise how the similarity of contemporary lexicons correlates with (historic) phylogeny and with the geographical closeness of speakers. As a way to make language evolution visual, we also provide the equivalent graph computed over data from historical linguistics, obtained from the ASJP database (Holman et al., 2011). Insights gained from these graphs may also help computational linguists predict the performance of automated tasks that involve some form of lexicon mapping (e.g. bilingual lexicon induction or machine translation) over specific language pairs.
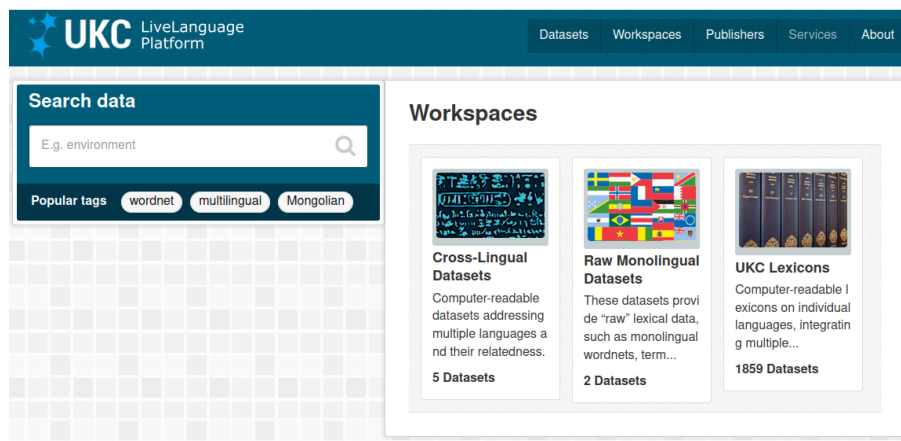
Figure 4: Cognate clusters for the concept of *fish*.

Figure 5: Detail of the *grandchild* subdomain as lexicalised by the Danish language.

Figure 6: Detail from the lexical similarity graph coloured according to language families.

Figure 7: The UKC LiveLanguage Data Catalogue.

5 The LiveLanguage Data Catalogue

As a complement to online exploration, we are also making available the contents of the UKC for computational applications. While an open, fine-grained, API-based access to the data is planned as future work, we are already in the process of publishing data for download through the *UKC LiveLanguage* data catalogue. The catalogue, accessible from the website through any of the numerous download links, provides access to the UKC data through multiple modalities. Accordingly, the structure of the catalogue, shown in Figure 7, consists of (1) cross-lingual datasets; (2) individual lexicons that aggregate all types of data about a single language; and (3) lexicon sets that also provide cross-lingual information. All datasets are published in full respect of the licensing constraints of their constituting resources; data that disallow redistribution are excluded from the catalogue.

Raw data on cross-lingual diversity. These datasets originate from projects on diversity and cover domain-specific lexical gaps, multilingual morphology, lexical similarity, and cognate relationships. Datasets are distributed in their original formats, with concepts linked to Princeton Word-Net 3.0 identifiers for interoperability with third-party data.

Individual lexicons. These datasets are produced as language-specific 'cross-sections' of the full UKC data. Their added value lies in the integration of multiple sources—words from wordnets and Wiktionary, language-specific morphological and lexico-semantic relationships, gaps—into a single formal representation. These datasets will be provided in multiple formats (under development),

including the ISO standard Lexical Markup Framework (LMF) format as well as OntoLex.

Lexicon sets. The notion of language diversity gains full significance *across* languages. Consequently, the development of an online service is underway to allow the simultaneous download of multiple concept-aligned lexicons as a single multilingual resource. The service will export multilingual data from the UKC database in real time. Such datasets will be directly exploitable in cross-lingual training and evaluation tasks.

6 Conclusions and Future Work

We see a huge research potential both in the creation and in the reuse of diversity-aware language resources. Exploiting diversity to improve state-of-the-art cross-lingual applications is a direction that we expect to gain importance and popularity in the community. The UKC database, website, and data catalogue aims to contribute to such efforts. All components of the system are going through a rapid evolution: the database contents in terms of language support, lexicon correctness and completeness, the data exploration and visualisation tools, new demonstrators, and APIs are continually being created and extended. We are also deploying regional instances of the UKC—e.g. for the Middle-East or the Indian subcontinent—where diversity among local languages and dialects is studied in fine detail by local communities of linguists.

162

References

Aryaman Arora, Adam Farris, Gopalakrishnan R, and Samopriya Basu. 2021. Bhāṣācitra: Visualising the dialect geography of South Asia. In *Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change 2021*, pages 51–57, Online. Association for Computational Linguistics.

Khuyagbaatar Batsuren, Gábor Bella, and Fausto Giunchiglia. 2019a. Cognet: A large-scale cognate database. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 3136–3145.

Khuyagbaatar Batsuren, Gábor Bella, and Fausto Giunchiglia. 2021a. A large and evolving cognate database. *Language Resources and Evaluation*, pages 1–25.

Khuyagbaatar Batsuren, Gábor Bella, and Fausto Giunchiglia. 2021b. Morphynet: a large multilingual database of derivational and inflectional morphology. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 39–48.

Khuyagbaatar Batsuren, Amarsanaa Ganbold, Altangerel Chagnaa, and Fausto Giunchiglia. 2019b. Building the mongolian wordnet. In *Proceedings of the 10th global WordNet conference*, pages 238–244.

Gábor Bella, Khuyagbaatar Batsuren, and Fausto Giunchiglia. 2021. A database and visualization of the similarity of contemporary lexicons. In *International Conference on Text, Speech, and Dialogue*, pages 95–104. Springer.

Gábor Bella, Fiona McNeill, Rody Gorman, Caoimhín Ó Donnaíle, Kirsty MacDonald, Yamini Chandrashekar, Abed Alhakim Freihat, and Fausto Giunchiglia. 2020. A major wordnet for a minority language: Scottish gaelic. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2812–2818.

Luisa Bentivogli, Pamela Forner, Bernardo Magnini, and Emanuele Pianta. 2004. Revising the wordnet domains hierarchy: semantics, coverage and balancing. In *Proceedings of the workshop on multilingual linguistic resources*, pages 94–101.

Pushpak Bhattacharyya. 2010. Indowordnet. In *In Proc. of LREC-10*. Citeseer.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362.

Lars Borin, Markus Forsberg, and Lennart Lönngren. 2008. The hunting of the blark–saldo, a freely available lexical database for swedish language technology. *Resourceful language technology. Festschrift in honor of Anna Sågvall Hein*, (7):21–32.

Johannes Dellert, Thora Daneyko, Alla Münch, Alina Ladygina, Armin Buch, Natalie Clarius, Ilja Grigorjew, Mohamed Balabel, Hizniye Isabella Boga, Zalina Baysarova, et al. 2020. Northeuralex: a wide-coverage lexical database of northern eurasia. *Language resources and evaluation*, 54(1):273–301.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. WALS Online. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Darja Fišer, Jernej Novak, and Tomaž Erjavec. 2012. slownet 3.0: development, extension and cleaning. In *Proceedings of 6th International Global Wordnet Conference (GWC 2012)*, pages 113–117.

Ruth Vatvedt Fjeld and Lars Nygaard. 2009. Nornet- a monolingual wordnet of modern norwegian. In *NODALIDA 2009 workshop: WordNets and other Lexical Semantic Resources-between Lexical Semantics, Lexicography, Terminology and Formal Ontologies*, volume 7, pages 13–16.

Amarsanaa Ganbold, Altangerel Chagnaa, and Gábor Bella. 2018. Using crowd agreement for wordnet localization. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Radovan Garabík. 2010. Slovak national corpus tools and resources. In *5th Workshop on Intelligent and Knowledge oriented Technologies*, page 2. Citeseer.

Radovan Garabík and Indrė Pileckytė. 2013. From multilingual dictionary to lithuanian wordnet. *Natural Language Processing, Corpus Linguistics, E-Learning*, pages 74–80.

Fausto Giunchiglia, Khuyagbaatar Batsuren, and Gabor Bella. 2017. Understanding and exploiting language diversity. In *IJCAI*, pages 4009–4017.

Fausto Giunchiglia, Khuyagbaatar Batsuren, and Abed Alhakim Freihat. 2018. One world–seven thousand languages. In *Proceedings 19th International Conference on Computational Linguistics and Intelligent Text Processing, CiCling2018, 18-24 March 2018*.

Aitor Gonzalez-Agirre, Egoitz Laparra, and German Rigau. 2012. Multilingual central repository version 3.0. In *LREC*, pages 2525–2529.

Marissa Griesel and Sonja Bosch. 2014. Taking stock of the african wordnet project: 5 years of development. In *Proceedings of the Seventh Global Wordnet Conference*, pages 148–153.

Maria Grigoriadou, Harry Kornilakis, Eleni Galiotou, Sofia Stamou, and Evangelos Papakitsos. 2004. The software infrastructure for the development and validation of the greek wordnet. *Romanian Journal of Information Science and Technology*, 7(1-2):89–105.

Eric W Holman, Cecil H Brown, Søren Wichmann, André Müller, Viveka Velupillai, Harald Hammarström, Sebastian Sauppe, Hagen Jung, Dik Bakker, Pamela Brown, et al. 2011. Automated dating of the world's language families based on lexical similarity. *Current Anthropology*, 52(6):841–875.

Chu-Ren Huang, Shu-Kai Hsieh, Jia-Fei Hong, Yun-Zhu Chen, I-Li Su, Yong-Xiang Chen, and Sheng-Wei Huang. 2010. Chinese wordnet: Design, implementation, and application of an infrastructure for cross-lingual knowledge processing. *Journal of Chinese Information Processing*, 24(2):14–23.

Hitoshi Isahara, Francis Bond, Kiyotaka Uchimoto, Masao Utiyama, and Kyoko Kanzaki. 2008. Development of the japanese wordnet. In *LREC*.

Mathieu Jacomy, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian. 2014. Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. *PloS one*, 9(6):e98679.

Ian Joo. 2021. The etymology of korean ssal 'uncooked grain' and pap 'cooked grain'. *Cahiers de Linguistique Asie Orientale*, 50(1):94 – 110.

Farhad Keyvan, Habib Borjian, Manuchehr Kasheff, and Christiane Fellbaum. 2007. Developing persianet: The persian wordnet. In *3rd Global wordnet conference. Citeseer*, pages 315–318. Citeseer.

Maria Koptjevskaja-Tamm, Ekaterina Rakhilina, and Martine Vanhove. 2015. The semantics of lexical typology. In *The Routledge handbook of semantics*, pages 450–470. Routledge.

Adrienne Lehrer. 1970. Notes on lexical gaps. *Journal of linguistics*, 6(2):257–261.

Krister Lindén and Lauri Carlson. 2010. Finnwordnet–finnish wordnet by translation. *LexicoNordica–Nordic Journal of Lexicography*, 17:119–140.

Marek Maziarz, Maciej Piasecki, and Stan Szpakowicz. 2012. Approaching plwordnet 2.0. In *Proceedings of 6th International Global Wordnet Conference, The Global WordNet Association*, pages 189–196.

George A Miller. 1998. *WordNet: An electronic lexical database*. MIT press.

Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Nurril Hirfana Bte Mohamed Noor, Suerya Sapuan, and Francis Bond. 2011. Creating the open wordnet bahasa. In *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation*, pages 255–264.

Noam Ordan and Shuly Wintner. 2007. Hebrew wordnet: a test case of aligning lexical databases across languages. *International Journal of Translation*, 19(1):39–58.

Petya Osenova and Kiril Simov. 2018. The data-driven bulgarian wordnet: Btbwn. *Cognitive Studies| Études cognitives*, (18).

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. Multiwordnet: developing an aligned multilingual database. In *First international conference on global WordNet*, pages 293–302.

Edoardo Maria Ponti, Helen O'horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*, 45(3):559–601.

Marten Postma, Emiel van Miltenburg, Roxane Segers, Anneleen Schoen, and Piek Vossen. 2016. Open dutch wordnet. In *Proceedings of the 8th Global WordNet Conference (GWC)*, pages 302–310.

Alexandre Rademaker, Valeria de Paiva, Fabricio Chalub, Livy Real, and Claudia Freitas. 2016. Introducing openwordnet-pt: a open portuguese wordnet for reasoning.

Ida Raffaelli, Marko Tadic, Božo Bekavac, and Željko Agic. 2008. Building croatian wordnet. In *Proceedings of GWC*, pages 349–360.

Yasser Regragui, Lahsen Abouenour, Fettoum Krieche, Karim Bouzoubaa, and Paolo Rosso. 2016. Arabic wordnet: New content and new applications. In *Proceedings of the 8th Global WordNet Conference (GWC)*, pages 333–341.

Christoph Rzymski, Tiago Tresoldi, Simon J Greenhill, Mei-Shin Wu, Nathanael E Schweikhard, Maria Koptjevskaja-Tamm, Volker Gast, Timotheus A Bodt, Abbie Hantgan, Gereon A Kaiping, et al. 2020. The database of cross-linguistic colexifications, reproducible analysis of cross-linguistic polysemies. *Scientific data*, 7(1):1–12.

Benoît Sagot and Darja Fišer. 2008. Construction d'un wordnet libre du français à partir de ressources multilingues. In *Actes de la 15ème conférence sur le Traitement Automatique des Langues Naturelles. Articles longs*, pages 171–180.

Sareewan Thoongsup, Thatsanee Charoenporn, Kergrit Robkop, Tan Sinthurahat, Chumpol Mokarat, Virach Sornlertlamvanich, and Hitoshi Isahara. 2009. Thai wordnet construction. In *Proceedings of the 7th Workshop on Asian Language Resources (ALR7)*, pages 139–144.

Dan Tufiş, Radu Ion, Luigi Bozianu, Alexandru Ceauşu, and Dan Ştefănescu. 2007. Romanian wordnet: Current state, new applications and prospects. In *Proceedings of the Fourth Global WordNet Conference (GWC 2008)*, pages 441–452.

PJTM Vossen. 1997. Eurowordnet: a multilingual database for information retrieval.

Shan Wang and Francis Bond. 2013. Building the chinese open wordnet (cow): Starting from core synsets. In *Proceedings of the 11th Workshop on Asian Language Resources*, pages 10–18.

Xinyi Wu, Zhenyao Wu, Hao Guo, Lili Ju, and Song Wang. 2021. Dannet: A one-stage domain adaptation network for unsupervised nighttime semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15769–15778.