Methods in Lexical Semantics
Language Maintenance and Low Resource Languages

Francis **Bond**

Department of Asian Studies,
Palacký University, Olomouc, Czechia

<bond@ieee.org>

LOT Winter School 2026

Faculty of Arts

Outline

Faculty of Arts

Roadmap

Faculty of Arts

Why FAIR Data?

- **Findable** — So that other researchers (and future you) can actually discover your data.
  If no one can find it, it might as well not exist.
- **Accessible** — So that once found, data can be retrieved easily and safely.
  Access should be possible even years later, with clear rules if restrictions apply.
- **Interoperable** — So that data from different projects can be combined or compared.
  Shared formats and vocabularies let computers and people understand each other.
- **Reusable** — So that data can be meaningfully used beyond its original purpose.
  Good documentation and clear licensing allow others to build on your work.

FAIR data makes research more transparent, verifiable, and sustainable

Faculty of Arts

Open licences and the Open Definition

- **Open data** means more than free access — it means legal permission to *use, modify, and share* without discrimination.
- The **Open Definition** (Open Knowledge Foundation) states that data is open if
  "anyone can freely access, use, modify, and share for any purpose"
  — subject only to requirements of attribution and share-alike.
- Common open licences:
  - **CC BY** – reuse with attribution.
  - **CC BY-SA** – reuse with attribution and same licence (share-alike).
  - **CC0 /Public Domain** – no restrictions.
  - **ODC BY/ODbL** – for databases; require attribution and share-alike.
- Choosing the right licence ensures that your data remains reusable and legally safe.

Openness is a design choice—licensing makes it possible, clarity makes it trustworthy.

Faculty of Arts

When not to use open licences

- Openness is a virtue, but not a universal rule.
- Some data should *not* be released under open licences because openness could cause harm.
- Examples include:
  - Personal or medical information that identifies individuals.
  - Cultural or linguistic materials shared under community protocols.
  - Locations of endangered species or sacred sites.
  - Data collected without full, informed consent for public reuse.
- In such cases, use restricted or tiered access, or licences that reflect community agreements.

Responsible openness means balancing transparency with care, consent, and context.

Faculty of Arts

Roadmap

Faculty of Arts

CARE: what and why

- **FAIR** focuses on making data easy to find and reuse.
- **CARE** ensures that openness respects people, communities, and context.
- Developed by the **Global Indigenous Data Alliance (GIDA)** (Carroll et al., 2020).
  - **Collective Benefit:** Who gains from sharing this data?
  - **Authority to Control:** Who decides how it can be used?
  - **Responsibility:** Are researchers accountable to those represented?
  - **Ethics:** How can openness coexist with respect and consent?

CARE asks: who benefits, who decides, and how can openness support justice?

Together, FAIR + CARE promote data that is both open and just.

Faculty of Arts

Towards green open research

- Each stage of the data lifecycle has environmental costs: storage, transfer, and computation.
- Good FAIR practice—clean metadata, smaller files, open formats—cuts energy use and prolongs data life.
- CARE principles promote community partnerships that reduce extractive, short-term projects.
- Reproducible workflows and open infrastructure avoid unnecessary duplication of effort and computation.

Practical takeaway

When you make data open, make it *efficiently* open: small, meaningful, reused, and responsibly hosted.

Sustainable openness means doing more with less—knowledge without waste.

Faculty of Arts

Roadmap

Faculty of Arts

Challenges in Building Wordnets for Low-Resource Languages

- Many languages lack lexical resources like dictionaries or corpora.
- In this case we have to build a foundational lexical database from scratch.
- Fieldwork and collaboration with native speakers are essential for data collection.
- Cultural and contextual differences make direct translations of concepts difficult.
- The lack of standardized orthographies adds complexity to digitization and documentation.

Faculty of Arts

Data Collection Challenges

- Workshops, interviews, and manual documentation are often necessary for word collection.
- Language consultants may be required for translating and interpreting lexical data.
- Typically language experts and computer experts are not the same people.
- Word meanings may vary by dialect or region, making consistent data collection difficult.
- Language communities with no history of standardisation may disagree as to what should be considered correct
- Time and cost constraints limit how much data can be collected. also true for well-resourced languages!

Faculty of Arts

Semantic Structure Challenges

- Identifying and organizing synsets in under-resourced languages can be difficult.
- Concepts may not map directly onto concepts in other languages, such as English.
- Cultural concepts and practices often require unique synsets that do not exist in larger languages.
- Expert knowledge of the language's lexical semantics is essential for accurate synset creation.

Technological Challenges

- Tools and resources for building wordnets, such as NLP software, are not readily available for low-resource languages.
- Manual annotation and digitization of handwritten data are time-consuming and error-prone.
- Often it is not just that you don't have, for example, a part-of-speech tagger, but that no-one had yet identified what parts of speech are appropriate
- Collaboration between linguists and native speakers can be logistically challenging without reliable power, internet or software tools.

Sociolinguistic and Cultural Challenges

- Some communities may resist linguistic documentation efforts due to concerns about language preservation.
- Native speakers may prioritize language revitalization over computational resources like wordnets.
- The process of documenting a language for a wordnet can introduce external cultural biases.
- ⋆ Collaboration with local communities is essential to ensure cultural respect and accuracy.
- ⋆ Ethical considerations are crucial when engaging with endangered language communities.

Faculty of Arts

Challenges in Data Validation and Accuracy

- Data collected from native speakers will contain inconsistencies or errors.
- Language experts are often needed to validate synsets and relations.
- Regular updates and revisions are necessary as languages evolve or more data is collected.
- Data verification is particularly challenging for languages with no formal linguistic documentation, and few or no collections of text.

Community Involvement

- Involving the language community is key to successful wordnet development.
- Native speakers contribute cultural and contextual insights critical to building accurate wordnets.
- Community workshops help ensure that the wordnet reflects the language as spoken by its speakers.
- Community members often play a key role in digitizing and validating the wordnet.
- Without community involvement, the wordnet may not reflect the real linguistic and cultural landscape.

Faculty of Arts

Wordnet is not designed to cover everything

- The original Princeton wordnet only covered content words: nouns, verbs, adjectives and adverbs in English
  - There was no need to cover every word class — plenty of other dictionaries do that
- But if a language has no dictionary, so the wordnet is going to be the **only** lexicon — then it needs to cover everything the community needs
  - Other parts of speech
  - Usage notes
  - Audio
  - Spelling variation
- And the data should be as accessible as possible
  - to the community
  - to the field linguists
  - to other researchers

Faculty of Arts

Balancing Preservation and Innovation

- Wordnets can support language preservation efforts by documenting lexical data.
- At the same time, they promote technological innovation, enabling computational uses of the language.
- There is often tension between focusing on language preservation versus creating resources for NLP applications.
- Balancing traditional language documentation with new technological tools is a key challenge.
- Wordnets offer a way to bridge the gap between preserving linguistic heritage and advancing technology.

Faculty of Arts

Roadmap

Faculty of Arts

Case Study 1: The Abui Wordnet

- Abui (ISO 639-3: abz, abui1241) is a Timor-AlorPantar (TAP) language (Kratochvíl, 2007)
- Spoken by about 40 thousand speakers in Central Alor
- We worked with the village of Takalelang on the northern coast.
- The Abui wordnet was developed as part of a fieldwork project documenting the language.
- Collected data is in the process of being (manually) digitized and structured into a wordnet format.

Faculty of Arts

Alor


Indonesia
International boundary
National capital
Railroad
Road
THAILAND
BURMA
LAOS
VIETNAM
CAMBODIA
PHILIPPINES
Andaman Sea
China Sea
Philippine Sea
South China Sea
Sulu Sea
Celebes Sea
Banda Sea
Arafura Sea
Timor Sea
Java Sea
Molucca Sea
PACIFIC OCEAN
INDIAN OCEAN
MALAYSIA
BRUNEI
SINGAPORE
Borneo
Kalimantan
Sumatra
GREATER SUNDA ISLANDS
LESSER SUNDA ISLANDS
New Guinea
PAPUA NEW GUINEA
EAST TIMOR
AUSTRALIA
PALAU
Alor-Pantar

Toolbox to wordnet

\lx pok	⇐ **lemma**
\ph ˈpɔk	⇐ **pronunciation**
\ps v.0 \pn kki	⇐ **part of speech**
\ge split \re broken \re smashed \re split \de split, burst, hatch, broken, smashed	⇐ **gloss, reversal, and definition (ENG)**
\gn pecah \rn retak \rn menetas \rn pecah \dn pecah, menetas, retak	⇐ **gloss, reversal, and definition (IND)**
\gr peca \rr menetas \rr peca \dr peca, menetas	⇐ **gloss, reversal, and definition (MLZ)**
\ref Poku.001 \xv Pingai nu hayei poku. \xe A plate fell down and broke. \xn Piring itu jatuh dan pecah. \xr Piring tu jatu peca.	⇐ **example sentence and translations**

- Uses the glosses to link to English, Malay and Indonesian wordnets
- Intersection in 3 language has an accuracy of 0.99, 2 languages around 0.5 and 1 language 0.35
  - Even though Malay and Indonesian are very similar!
- Data made available at `https://github.com/fanacek/abuiwn`
- This wordnet was built using other wordnets

(Kratochvil and Morgado da Costa, 2022)

Faculty of Arts

Rapid Word Construcion

- But we wanted more words!
- Three RWC workshops were conducted with over 80 participants across multiple days.
- Workshops focused on gathering words within specific semantic domains.
- Participants included native speakers from different parts of the community, ensuring dialectal coverage.
- Words were first handwritten, then digitized and annotated for semantic relations.
- SIL domains were mapped to wordnet concepts, again using translation overlap (bin Mohd Rosman et al., 2014; Morgado Da Costa et al., 2023)
  - because the granularity is very different they do not often match one-to-one
  - so there is more manual work to be done

Faculty of Arts

It was a blast

SIL semantic domain for water (1.3)

1.3 Water

Use this domain for general words referring to water.

Related domains: 6.6.7 Working with water
7.2.4.2 Travel by water

Louw Nida Codes: 2D Water

What general words refer to water?

water, H2O, moisture

What words describe something that belongs to the water or is found in water?

watery, aquatic, amphibious

What words describe something that water cannot pass through?

waterproof, watertight

- » 1.3.1 Bodies of water
- » 1.3.2 Movement of water
- » 1.3.3 Wet
- » 1.3.4 Be in water
- » 1.3.5 Solutions of water
- » 1.3.6 Water quality

‹ 1.2.3.3 Gas up 1.3.1 Bodies of water ›

```
<rt class="CmDomainQ" guid="6fa93eab-71e0-4880-9a78-0b2a81882800" ownerguid="603
974-a005-4567-82e9-7aaeff894ab0">
<ExampleWords>
<AUni ws="en">water, H2O, moisture</AUni>
<AUni ws="es">agua, H2O, humedad, preciado líquido</AUni>
<AUni ws="fa">آب، رطوبت، نم، خیس</AUni>
<AUni ws="fr">eau, H2O, humidité</AUni>
<AUni ws="hi">पानी, H2O, नमी</AUni>
<AUni ws="id">air, H2O, embun</AUni>
<AUni ws="ne">पानी, जल, नीर, तरल</AUni>
<AUni ws="pt">água, H2O, humidade</AUni>
<AUni ws="ur">پانی، آب</AUni>
<AUni ws="zh-CN">水、H2O</AUni>
</ExampleWords>
<Question>
<AUni ws="bn">(১) পানি বোঝাতে সাধারনত কি কি শব্দ ব্যবহার করা হয়?</AUni>
<AUni ws="en">(1) What general words refer to water?</AUni>
<AUni ws="es">(1) ¿Cómo se le llama generalmente al agua?</AUni>
<AUni ws="fa">چه واژگانی به آب مربوط می‌شوند؟</AUni>
<AUni ws="fr">(1) Quels sont les termes génériques qui désignent l'eau_?</AUni>
<AUni ws="hi">(1) पानी?</AUni>
<AUni ws="id">(1) Kata-kata umum apa yang digunakan untuk menyebut air?</AUni>
<AUni ws="ne">(१) साधारण कुन-कुन शब्दहरूले पानी जनाउँछ?</AUni>
<AUni ws="pt">Que palavras gerais referem à água?</AUni>
<AUni ws="ru">(1) Какие основные слова относятся к воде?</AUni>
<AUni ws="th">คำทั่วไปคำใดที่หมายถึงน้ำ? น้ำ, ความชุ่มชื้น, เหลว</AUni>
<AUni ws="ur">پانی کے لئے عام کونسے الفاظ استعمال کئے جاتے ہیں؟</AUni>
<AUni ws="zh-CN">通常说到水，你会怎么说？</AUni>
</Question>
</rt>
```

Faculty of Arts

Challenges in Building the Abui Wordnet

- Data entry was the biggest bottleneck
  - Only a native speaker can really digitize the data
  - There are only a handful who could
  - They often have other things to do
- Computational linguist time is also a bottleneck
  - Resource creation is rarely well-funded
  - It takes second place to other tasks
- Orthographic variation is also a problem
  - The orthography is being refined as we record more data
  - It is hard to update older data

Faculty of Arts

Outcomes of the Abui Wordnet Project

- The Abui wordnet documented over 1,400 synsets and 3,600 senses
- A new version with 2,500 synsets is on its way
- It serves as a lexical resource for both linguistic research and the local community.
- The project supported ongoing language documentation and preservation efforts.
- Native speakers have been deeply involved in the project and are training in linguistics
- The Abui wordnet was incorporated into the Open Multilingual Wordnet (OMW) project — you can even find it in hugging face!
- We are currently working on annotating a text: *Bukuuting bikaat-bikaat* “The Speckled Band’, which we translated from Indonesian
  for this we need a lemmatizer, …

Faculty of Arts

Case Study 2: The Kristang Wordnet

- Kristang is a critically endangered creole language
- Spoken mainly by Portuguese-Eurasian communities in Malacca and Singapore.
- There are no more than a few thousand speakers, with more in Malacca than Singapore.
- Kristang is a originally derived from Malay and Portuguese
  - Vocabulary is largely from Portuguese
  - Grammar is very close to Malay
  - Influenced also by Dutch, English and also other Portuguese creoles in Africa, India, South East Asia and China

Faculty of Arts

Portuguese Creoles in Asia

Language Revitalization and the Kristang Wordnet

- The Kristang wordnet is part of a language revitalization effort *Kodrah Kristang* “Awaken Kristang”
- The goal is to support the dwindling speaker base (Morgado da Costa, 2020).
- The wordnet aims to preserve Kristang’s lexicon and support its transmission to younger generations.
- The project is led by local community members, often linguistic students.
- Classes and workshops are held to teach Kristang using materials derived from the wordnet.
- Workshops are also held to add new words to the wordnet
- Community involvement is central, with local speakers helping to curate and validate data, efforts are supported by partnerships with linguistic organizations and local governments.

Faculty of Arts

Kodrah Kristang Class

Challenges in Building the Kristang Wordnet

- Kristang's limited written tradition required reliance on a wide variety of language resources.
  - paper dictionaries and word lists, including personal collections
  - linguistic publications with wordlists or glossed text
  - language documentation work (including *Kodrah Kristang* and a course at NUS on Field Methods in Linguistics)
  - new words and senses produced by `Jardinggu` "langarden", the Kristang lexical incubation project
- The small speaker base makes data collection and validation challenging.

Faculty of Arts

Innovation in the Kristang Wordnet Project

- The project took advantage of existing work in the open multilingual wordnet to also allow an extended inventory of parts of speech: pronouns, classifiers, exclamatives and so forth (Seah and Bond, 2014; Morgado da Costa and Bond, 2016)
- It is also important to make the lexical data available even when not fully integrated
  - In order to be useful to the community the Open Kristang Wordnet has two layers
  - The core wordnet with synsets fully linked to the OMW
  - The extended wordnet that just has bilingual Kristang-English pairs

Faculty of Arts

Significance of the Kristang Wordnet

- The wordnet supports efforts to reverse language shift in the Kristang-speaking community.
- It highlights the importance of collaboration between linguists and language communities.
- The Kristang wordnet can serve as a model for revitalization efforts in other endangered languages.
- It demonstrates how lexical resources can play a role in both preservation and education.
- The project shows the potential for wordnets in revitalizing endangered creoles and minority languages.

Faculty of Arts

Case Study 3: The Cantonese Wordnet

- Cantonese is spoken by millions, but its written tradition is limited compared to Mandarin.
- The Cantonese wordnet project aims to provide a lexical resource for this major Chinese variety.
- The Cantonese wordnet includes everyday Cantonese vocabulary, colloquialisms, and slang.
- It currently has senses, examples, a small sense-tagged corpus (Sio and Costa, 2019; Sio and Morgado da Costa, 2022)
- It also has parts-of-speech not in the original wordnet: classifiers and aspect markers

Faculty of Arts

Where is Cantonese (and other Chineses) spoken?

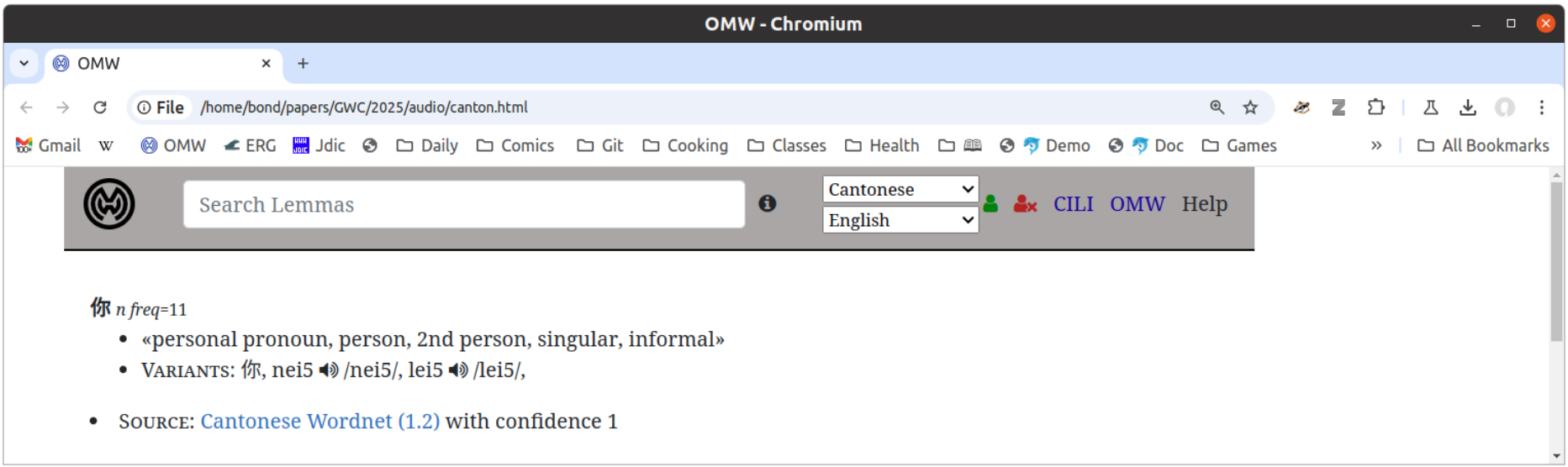
From Zhenxing et al. (2012).

Challenges in Building the Cantonese Wordnet

- Cantonese has a strong oral tradition but not a standardized written forms.
- Because of this there is quite a bit of variation
- As Cantonese is a spoken variety, and there is some variation in pronunciation, we decided to add audio to the wordnet
  - This takes advantage of the extensions to the wordnet format from 2020
  - As far as we know we are the first people to add sound
  - We recorded some data ourselves
  - We harvested some data from Wikicommons
    - Many entries were Mandarin with Cantonese pronunciation, not Cantonese
    - It is difficult to distinguish them
    - But essential to do so
- It was sometimes hard to distinguish words from phrases
  - Written Chinese does not show word boundaries
  - Linguists disagree on what should be a word

Faculty of Arts

The OMW interface showing audio



- This shows a word with two pronunciations
  - nei5 is the standard pronunciation
  - lei5 is the 'lazy' pronunciation (Chen, 2018)

Faculty of Arts

Outcomes of the Cantonese Wordnet Project

- The Cantonese wordnet is a carefully curated resource
  - 6,200 concepts
  - 17,350 senses
  - 2,138 audio examples, covering 2,859 senses
- The project aids in preserving Cantonese as a separate linguistic entity from Mandarin.
- The wordnet has been incorporated into multilingual wordnets, supporting cross-linguistic research.
- It serves as a basis for future research on Cantonese language technology development.

Faculty of Arts

Significance of the Cantonese Wordnet

- The Cantonese wordnet serves as a reference for the vocabulary of the language.
- It supports the development of linguistic resources for non-Mandarin Chinese varieties.
- The wordnet helps capture the cultural and linguistic identity of Cantonese speakers.
- It contributes to the broader effort of documenting and preserving regional Chinese languages.
- The project highlights the importance of creating linguistic resources for spoken varieties.

Faculty of Arts

Roadmap

Faculty of Arts

Advantages of Wordnets for Low-Resource Languages

- Wordnets help document and preserve linguistic data for endangered or under-documented languages.
- They enable cross-linguistic comparison by aligning synsets with wordnets of other languages.
- Wordnets provide a structured, searchable resource that benefits researchers and language communities.
- They can be used to support language revitalization efforts and to create educational materials.
  - Having an online presence can boost the social-status of a language
- Wordnets help build other wordnets
  - More languages allows better bootstrapping
  - Phenomena covered in one language make it easier for the next
  - The more data there is, the better the descriptions become

Faculty of Arts

Cultural and Linguistic Preservation

- Wordnets should record and organize culturally significant concepts that may not exist in other languages.
  - This has not been done as much as it should
  - Adding and describing new concepts is the next great challenge!
- They create a permanent record of a language's lexicon, supporting long-term preservation.
- Wordnets reflect the unique cultural and cognitive world of speakers, documenting traditional knowledge.
- In endangered languages, wordnets can capture the lexical heritage before it disappears.
- They facilitate the transmission of traditional vocabulary to younger generations.

Faculty of Arts

Technological Benefits

- Wordnets support the development of language technology for low-resource languages.
- They provide a foundation for tools like speech recognition, machine translation, and sentiment analysis.
- Wordnets enable the development of intelligent systems that can understand semantic relationships in texts.
- They help create more inclusive digital resources, bringing under-represented languages online.
- Open-source platforms like the Open Multilingual Wordnet (OMW) facilitate easy integration into existing projects.

Faculty of Arts

Educational and Revitalization Efforts

- Wordnets can be used to create language learning materials and dictionaries for education.
- They provide an organized resource for teaching both linguistic structure and vocabulary.
- Revitalization programs benefit from wordnets by using them to foster literacy and language proficiency.
- Wordnets allow the community to access and engage with their language in new technological formats.
- They offer a digital resource for the continued learning and use of the language in modern contexts.

Faculty of Arts

Advantages in Linguistic Research

- Wordnets allow linguists to analyze and compare the lexical structure of different languages.
- They provide insights into how languages encode meanings and semantic relationships.
- Multilingual wordnets facilitate cross-linguistic research on polysemy, metaphor, and lexical relations.
- Wordnets serve as an important resource in typological studies and language contact research.
  - For example we used the Moroccan Arabic wordnet to investigate shared vocabulary with Standard Arabic, French and Italian (Mrini and Bond, 2018)
- They enable the study of cognitive aspects of language, such as categorization and conceptualization.

Faculty of Arts

Roadmap

Faculty of Arts

Thanks and disclaimer

- An earlier version of this lecture was given at the Teanga Project.
- Thanks to all the many people who have worked on these resources, especially Luís Morgado da Costa, František Kratochvíl and Joanna Ut-Seong Sio.
- I have talked about projects my lab has been involved in, as I know them best, but there are many other wordnets for LRL
  - Amharic, Kurdish, Mansi, Moroccan Arabic, Sardinian, ASL, Uzbek, Welsh, …
- There are also wordnets for ancient languages
  - Ancient Greek, Coptic, Latin, Qin Chinese, Sanskrit…

References I

Muhammad Zulhelmy bin Mohd Rosman, Francis Bond, and František Kratochvíl. 2014. Bringing together over- and under-represented languages: Linking wordnet to the SIL semantic domains. In *Proceedings of the 7th Global WordNet Conference (GWC 2014)*. Tartu. 40–48.

Stephanie Russo Carroll, Ibrahim Garba, Olga Figueroa-Rodríguez, Jarita Holbrook, Raymond Lovett, Sylvia Materechera, Mark Parsons, Kay Raseroka, Desi Rodriguez-Lonebear, Robyn Rowe, Rebecca Sara, John Walker, Jane Anderson, and Maui Hudson. 2020. The care principles for indigenous data governance. URL https://www.gida-global.org/care, global Indigenous Data Alliance.

Katherine Hoi Ying Chen. 2018. Ideologies of language standardization: The case of Cantonese in Hong Kong. In *The Oxford Handbook of Language Policy and Planning*. OUP.

Faculty of Arts

References II

Frantisek Kratochvil and Luís Morgado da Costa. 2022. Abui Wordnet: Using a toolbox dictionary to develop a wordnet for a low-resource language. In Oleg Serikov, Ekaterina Voloshina, Anna Postnikova, Elena Klyachko, Ekaterina Neminova, Ekaterina Vylomova, Tatiana Shavrina, Eric Le Ferrand, Valentin Malykh, Francis Tyers, Timofey Arkhangelskiy, Vladislav Mikhailov, and Alena Fenogenova, editors, *Proceedings of the first workshop on NLP applications to field linguistics*, pages 54–63. International Conference on Computational Linguistics, Gyeongju, Republic of Korea. URL https://aclanthology.org/2022.fieldmatters-1.7.

František Kratochvíl. 2007. *A Grammar of Abui: A Papuan Language of Indonesia*. LOT, Utrecht.

Faculty of Arts

References III

Luís Morgado da Costa. 2020. Pinchah kristang: A dictionary of kristang. In Ilan Kernerman, Simon Krek, John P. McCrae, Jorge Gracia, Sina Ahmadi, and Besim Kabashi, editors, *Proceedings of the 2020 Globalex Workshop on Linked Lexicography*, pages 37–44. European Language Resources Association, Marseille, France. URL https://aclanthology.org/2020.globalex-1.7.

Luís Morgado da Costa and Francis Bond. 2016. Wow! what a useful extension to wordnet! In *10th International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož.

Luís Morgado Da Costa, František Kratochvíl, George Saad, Benidiktus Delpada, Daniel Simon Lanma, Francis Bond, Natálie Wolfová, and A.I. Blake. 2023. Linking SIL semantic domains to wordnet and expanding the Abui wordnet through rapid words collection methodology. In *12th International Global Wordnet*

Faculty of Arts

References IV

Conference (GWC2023). URL https://aclanthology.org/2023.gwc-1.38/.

Khalil Mrini and Francis Bond. 2018. Putting figures on influences on Moroccan Darija from Arabic, French and Spanish using the wordnet. In Francis Bond, Takayuki Kuribayashi, Piek Vossen, and Christiane Fellbaum, editors, *Proceedings of the 9th Global Wordnet Conference (GWC 2018)*. Global Wordnet Asssociation, Singapore.

Yu Jie Seah and Francis Bond. 2014. Annotation of pronouns in a multilingual corpus of Mandarin Chinese, English and Japanese. In *10th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*. Reykjavik.

References V

Joanna Ut-Seong Sio and Luis Morgado Da Costa. 2019. Building the Cantonese Wordnet. In Piek Vossen and Christiane Fellbaum, editors, *Proceedings of the 10th Global Wordnet Conference*, pages 206–215. Global Wordnet Association, Wroclaw, Poland. URL https://aclanthology.org/2019.gwc-1.26.

Joanna Ut-Seong Sio and Luis Morgado da Costa. 2022. Enriching linguistic representation in the cantonese wordnet and building the new cantonese wordnet corpus. In *Proceedings of the 13th Conference on Language Resources and Evaluation*. European Language Resources Association (ELRA), Marseille, France.

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, et al. 2016. The fair guiding principles for scientific data management and stewardship. *Scientific Data*, 3(160018). URL https://doi.org/10.1038/sdata.2016.18.

Faculty of Arts

References VI

Zhang Zhenxing, Xiong Zhenghui, Huang Xing, Dao Bu, and Benjamin K. T'sou, editors. 2012. 中国语言地图集 [Atlas of Chinese Languages]. Commercial Press, second edition. 2 volumes.