# Is this Enough? An evaluation of the Malayalam Wordnet

**Nandu Chandran Nair, Maria-Chiara Giangregorio, Fausto Giunchiglia**
University of Trento
nandu.chandrannair@unitn.it,mariegiangregorio@protonmail.com
fausto.giunchiglia@unitn.it

## Abstract

The quality of a product is the degree to which a product meets the Customer's expectations, which must also be valid in the case of lexical-semantic resources. Conducting a periodic evaluation of resources is essential to ensure that they meet a native speaker's expectations and are free from errors. This paper defines the possible errors that a lexical-semantic resource can contain, how they may impact downstream applications and explains the steps applied to evaluate and quantify the quality of Malayalam WordNet. Malayalam is one of the classical languages of India. We propose an approach allowing to subset the part of the WordNet tied to the lowest quality scores. We aim to work on this subset in a crowdsourcing context to improve the quality of the resource.

## 1 Introduction

Internet is composed of machine-readable lexical or lexical semantic resources (Herring, 2008). However, only a small percentage of languages have these resources developed in full scale (Herring, 2008). Hence the majority of languages are classified as under-resourced (Besacier et al., 2014). In our study, we consider a language to be under-resourced if it has the following characteristics: lack of computer-readable resources, lack of linguist experts, and limited usage on the internet (Krauwer, 2003). The different large scale availability of language resources results in a digital language divide (Warschauer, 2002). Figure 1 shows the visual representation of the study's findings on language-wise Internet users in 2021 and was conducted by KPMG India. We notice that usage of Hindi is greater than Malayalam. Bringing more languages online may ultimately be an exercise in cultural preservation, rather than utility. The probability of retrieving a relevant result using an under-resourced language is comparatively lower than with languages falling in the opposite category (Wheeler and Dillahunt, 2018). To diminish the divide, we either need new methodologies that allow for the creation of language resources or the creation of a platform supporting the continuous development of such resources (Chakravarthi et al., 2018, 2019, 2020a; Chakravarthi, 2020).

A person communicates their need for knowledge through words with a precise meaning, however from a machine point of view meaning( as well as specific acceptations) do not automatically come in conjunction with the word. Lexical-semantic resources come into play to bridge this gap, associating meanings to words. The most commonly used lexical-semantic resource is the Princeton WordNet(PWN) (Miller, 1998). PWN is a lexical database for English that is organized around synsets. The popularity of PWN has spread in the language community and has lead to the generation of similar resources. One such attempt was made in India, a Country both rich in culture and languages. IndoWordNet(IWN) is first multilingual WordNet for Indian languages developed through the joint efforts of different reputed universities across the country. IWN developed intending to sketch India's cultures in length and breadth by including 18 languages out of 22 official languages (Dash et al., 2017). All the languages in the IWN does not have the same synset IDs as PWN. For each synset in IWN, there is a corresponding synset in English that represent the same concept. In the paper, (Nair et al., 2019) authors extended this idea to align the IWN with PWN with the help of a large-scale multilingual resource called the Universal Knowledge Core(UKC) (Tawfik et al., 2014). The



*Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 100–108

UKC has a similar structure of PWN and is designed as a multilayered ontology that has a language-independent semantic layer called the Concept Core (CC) and a language-specific lexico-semantic layer called the Language Core (LC). Hence, in the UKC, the meaning of the words are represented not only by the synsets but also using lexical concepts (Giunchiglia et al., 2018). Many works have been conducted under the UKC. Recent one is (Bella et al., 2020), where authors explores the unique grammatical properties of natural language text and perform experiments on tokenisation, part of speech tagging and named entity recognition over real-world structured data. Another important one is (Batsuren et al., 2019), which introduces a large-scale lexical database that provides words of common origin and meaning across languages.

Before enriching an existing resource, ensuring the quality and finding hidden errors is recommended in order to avoid producing a low-quality extensive resource. Inadequate quality resources are not suitable to train artificial intelligence tools (Chakravarthi et al., 2020b). One of the methods to ensure the quality of a resource is getting it validated using linguistic experts. Hiring linguists are expensive, or for some languages it is difficult. And another drawback is that contribution from one expert could be biased (Bonvillain, 2019). Hence we need faster and cheaper ways to estimate the quality of a WordNet. Also, the main question here is what are the factors that define the quality of a WordNet. As per the literature, a WordNet development team usually does structural check and whether the WordNet can perform applications like information retrieval (Baeza-Yates et al., 2015) and question-answering (Moldovan and Rus, 2001).

In this paper, we classify possible errors that affect the quality of the WordNet: schema errors and semantic errors. We list five schema errors from the IWN. We estimate semantic error by computing cosine similarity(Rahutomo et al., 2012) between two WordNets and within the WordNet using a pre-trained machine learning model(Zhou, 2016) for Indian languages. Using our approach, we generate a candidate set of synsets with very high semantic error that will be verified by native speakers in a crowdsourcing application(Brabham, 2013).

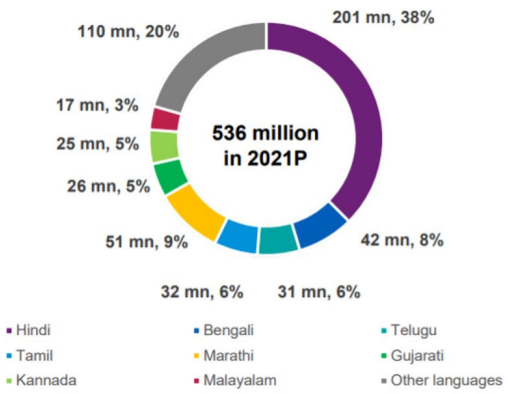



Figure 1: Language wise Internet users based on the study conducted by KPMG

This paper is structured as follows: Section 2 covers details about the language Malayalam and Malayam Wordnet (MWN), section 3 defines the quality of the WordNet and lists out the possible errors in WordNet, section 4 covers the procedure we used to find and establish the quality of the MWN, section 5 lists results and finally section 6 provides conclusions and direction for further work.

## 2 Malayalam Language

Malayalam(Asher, 2013) is a Dravidian language spoken in the Indian state of Kerala and union territories of Lakshadweep and Puducherry. Malayalam evolved from Tamil language after 16 century CE by Father of Malayalam Thunchaththu Ramanujan Ezhuthachan. In the early 21st century, Malayalam was spoken by around 38 million people all over the world (Asher, 2013). In the past, media like television, newspaper, short stories, novels, etc. helped with the diffusion and prospering of Malayalam language and literature(Palakeel, 1996). Nowadays smartphones, laptops, and Internet usage have become an important part of people's lives. The use of Malayalam language through these platforms has helped people to communicate, learn and express. Introduction of 'Unicode' technology in the Malayalam language helped correct spelling mistakes, grammatical errors, etc (Sooraj et al., 2018) (Santhosh et al., 2002), which was initially possible only in the English language. The stories and articles were in the regional language, though they used English words for dif-

ferent technical terms(Sasidharan et al., 2009). For example, the Malayalam word for mobile phones is "ദൂരഭാഷണശ്രവണസഹായി" (Dhoora bashana sravana sahayi), which was never used in any of the articles/blogs/news videos. The word 'Switch' in Malayalam was "വൈദ്യുത ആഗമന നിയന്ത്രണ യന്ത്രം" (Vaidhyutha agamana niyanthrana yanthram). But due to its complexity, these words were never used. The Malayalam language lost lots of words due to this reason(Prasanth, 2016). The Malayalam vocabulary is rich in terms from specific domains, with the notable example of terms reflecting family relations(George, 1972). In Kerala, where Malayalam is mother tongue, family relations are essential and there are particular names for each type of relation. A notable example of this richness is represented by the English word *cousin*: "the child of your aunt or uncle". In Malayalam the term used to denote a *cousin* will change depending on age, gender, whether the relation is from the paternal or maternal side. Sometimes the background of the family is also a factor influencing the selection of a different term. There is no such thing as a direct translation of the word *cousin* in Malayalam.

The Indian University, Amrita Vishwa Vidyapeetham, took part in the development of MWN (Rajendran and Soman, 2017) in 2011 as part of the project entitled "Development of Dravidian wordnet: an integrated wordnet for Telugu, Tamil, Kannada and Malayalam" which later integrated into IWN along with other Indian languages. The development of MWN is motivated by PWN. The initial idea of PWN was to aid the search in dictionaries, conceptually rather than alphabetically. Based on inspiration coming from current psycholinguistics (Foss, 1972) theories of human memory and its working, PWN organizes English *nouns*, *verbs*,*adverbs*, and *adjectives* into *synonym* sets or *synset*, each representing one underlying lexical concept. Synonyms are defined as words holding the same meaning in the same language (Miller, 1998). A set of synonymous words is called synset. For instance, "head" and "caput" are the synsets with the meaning "the front or the upper part of the body of animals bearing the face and brains". In PWN, synsets are associated with a *gloss*. A Gloss defines the meaning of a synset, the concept associated with the synset. MWN uses the same structure and lists the following entries for each concept: synset ID, synsets, gloss and example sentence. Synset ids used in MWN are different from PWN as the expand approach was used to develop MWN. In the expand approach, Hindi synsets are translated into Malayalam, this ensures concepts which are culturally relevant to India are retained as part of MWN. In our study, we are using source files of MWN we received from Amrita Vishwa Vidyapeetham. There are 30139 synsets available in MWN. In which, 20071 synsets are nouns, 3311 synsets are verbs, 501 synsets are adverbs, and 6256 synsets are adjectives.

## 3 Defining Quality

Within the translation industry, three terms are used somewhat interchangeably to refer to quality-related initiatives: quality assessment(House, 2014), quality assurance(Vîlceanu, 2017) and quality control(Ibba and Söll, 1999).

- Quality assessment or quality evaluation is the measurement of the extent to which a product complies with quality specifications(Lommel et al., 2014).
- Quality assurance refers to ways of preventing mistakes or defects in manufactured products and avoiding problems when delivering solutions or services to customers. Quality assurance relies on continual assessment of quality(Lommel et al., 2014).
- Quality control is the process of checking whether manufactured products meet stated quality specifications. While quality assurance relates to how a process is performed or how a product is made, quality control is tied more to the inspection aspect of quality management(Lommel et al., 2014).

Based on the definition, quality control requires specifications(Gouadec, 2010) which can then be used to assess whether a resource meets quality thresholds. In our work, we are performing quality control of the MWN. We are defining the minimum specifications in order

to evaluate the WordNet and assess whether it is a good quality resource.

Multidimensional Quality Metrics (MQM)(Lommel et al., 2014) provides a framework for describing and defining quality metrics used to assess the quality of translated texts and to identify specific issues in those texts. The scope of this framework is limited to quality assessment of translated content. It does not apply to the evaluation of the translation process or projects. MQM defines the quality as adherence of the text to appropriate specifications. As per MQM, specifications are a description of requirements for the translation. In our work, we have been inspired to adapt ideas used as part of MQM for defining quality specifications of the WordNet. In MQM, a target document quality score is defined as

$$TQ = 100 - TP + SP$$

TQ being the quality score, TP the penalties for target content, and SP the penalty tied to source content. We then define a resource as being of good quality if the resource is free from errors.

We define two categories of errors in WordNet: schema errors and semantic errors.

### 3.1 Schema Errors

Schema errors occur when there is a problem with the file's structure or order, or an invalid character is included. There are five observed schema errors. They are:

- Category A: Empty gloss and/or synset and/or example sentence
- Category B: Poorly defined gloss
- Category C: Unknown characters in the fields
- Category D: Poorly defined part of speech
- Category E: Duplicate gloss

We consider a record as having unique synset id, gloss, example sentence and synset. Category A is the number of records with empty gloss and/or empty example sentence and/or empty synset(Figure 2). In some WordNets, example sentences are optional, so we will do a manual check in the beginning to verify metadata. Category B represents the number of records having a poorly defined gloss (Figure 3). Poorly defined glosses are very general, and computing this value is a bit challenging. Hence, we consider the number of records that have the same gloss as the corresponding synset. Category C is the number of records that have unknown characters in any of the fields (Figure 4). Category D is very rare but possible, including records that have poorly defined parts of speech (Figure 5). Four parts of speech are included in MWN: noun, adjective, verb and adverb. Due to errors during the development phase, these values may have spelling mistakes. This eventually affects the quality of the WordNet. Category E includes the number of records having the same gloss (Figure 6). For these records, synset id, example sentence and synsets are different, gloss is the same.

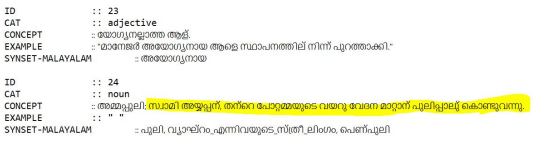



Figure 2: Example for Empty example sentence

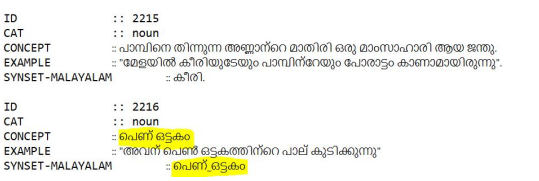



Figure 3: Example for poorly defined gloss

### 3.2 Semantic Errors

Meanings can be represented in many languages if the corresponding concept is available across all languages that are taken into account (Mercer, 2002). Meaning is defined as what is meant by a word, text, concept, or action. For example, the word "rain" means "water falling in drops from vapor condensed in the atmosphere" and the same meaning is represented in Italian as "pioggia", in Dutch as "regen". It is also worth keeping in mind that the word "rain" has two additional possible meanings, available depending on the context the word is used in.

Semantic errors are estimated when comparing across WordNets; checking how good conceptual alignment between the languages

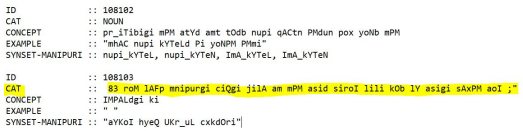



Figure 4: Example for poorly defined pos

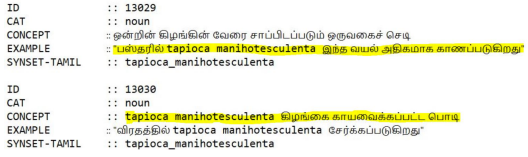



Figure 5: Example for unknown characters in gloss

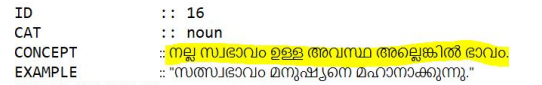
is. In our work, we are using PWN and HWN to find the semantic error in MWN. Semantic errors are caused by factors like cultural differences, linguistic differences, and mistakes made by humans.

Figure 7 listing out some examples from these category and we have taken these values from IWN. Figures 7a, 7c, 7d and 7e are synsets specific to the language and culture. Finding out this candidate set could help the researchers to understand the diversity and similarity (Giunchiglia et al., 2017) across languages.

## 4 Evaluating MWN

In this section, we evaluate schema errors and semantic errors for the MWN. We estimated the schema errors of the resource as listed in the table 1. There are three categories of schema errors presented in the MWN; category A, B and E. 181 records have empty example sentence, 1789 records have empty synsets and 2374 records have poorly defined glosses. And then we found 925 records with duplicate gloss.

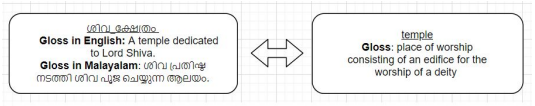
We sampled synsets common to English,



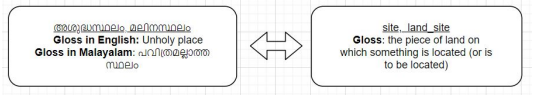

(a) For synset id 16


(b) For synset id 21858

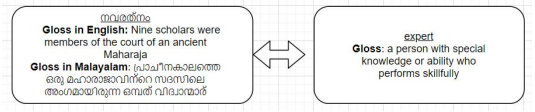
Figure 6: Example for duplicate gloss



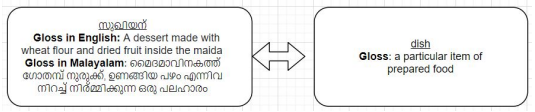

(a) Synset “temple” vs “ശിവ _ക്ഷേത്രം”(shiva_kshetram)



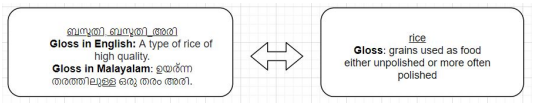

(b) Synset “site” vs “അശുദ്ധസ്ഥലം”(ashudhasthalam)




(c) Synset “expert” vs “നവരത്നം” (navarathnam)




(d) Synset “dish” vs “സുഖിയന്”(sukhiyan




(e) Synset “rice” vs “ബസ്മതി _അരി”(basmati _ari)

Figure 7: Examples for semantic errors

Hindi and Malayalam WordNets; this yielded 18938 synsets. We then computed cosine similarity in order to identify synsets tied to poor semantic alignment and through this be able to identify semantic errors. We propose a cosine similarity score based classification parsed into three classes: low alignment for scores lower than 0.4; moderate for scores between 0.4 and 0.7; strong for scores greater than 0.7. We have used the threshold of minimum 0.4 inspired by the work by (Khodak et al., 2017). In (Khodak et al., 2017) authors automatically generate WordNet data by using a machine-readable dictionary. Based on our classification, we will consider the sets of synsets having cosine similarity scores below 0.4 as requiring attention.

We used two approaches to measure semantic errors: we computed cosine similarity(semantic alignment) between two languages; we then calculated semantic alignment within the language. Algorithm 1 shows the steps followed in computing the semantic similarity between English gloss and Malayalam gloss. We have used the same steps to calculate the semantic similarity between English word sense and Malayalam word sense. And repeat the above steps also for language Hindi. Algorithm 2

| Type of schema error | No.synsetids |
|---|---|
| Empty example sentence | 181 |
| Empty synsets | 1789 |
| Poorly defined gloss | 2374 |
| Duplicate gloss | 925 |

Table 1: Schema errors in Malayalam lexical-semantic resource

shows the steps followed in computing the semantic similarity between a word sense in the synset and the gloss of MWN.

A pre-trained machine learning model was used in order to compute semantic alignment and cosine similarity. There are a number of pre-trained models available for similarity checking. Multilingual BERT(mBERT)(Pires et al., 2019), XLM-RoBERTa(Conneau et al., 2019) and Sentence transformers(Reimers and Gurevych, 2019) are only few of them. Multilingual BERT(mBERT) and XLM-RoBERTa have been known to produce less than ideal sentence representations when deployed out-of-box. They would additionally pose the problem of coming with non- aligned vector spaces across languages. As a result, sentences with the same semantic content, expressed with different languages, would be mapped to different vector spaces. As part of our work we used sentence transformers: a Python framework for state of the art sentence and text embeddings that can be used to compute sentence/text embeddings for more than 100 languages. These embeddings can be compared through cosine similarity, allowing for the identification of sentences with similar meaning. The framework is based on PyTorch(Paszke et al., 2019), Transformers and offers a large collection of pre-trained models tuned for various tasks. We have used the pre-trained multilingual model *stsb-xlm-r-multilingual* and aligned vector spaces allowing for similar inputs across languages to be mapped close within the same vector space. XLM-R supports 100 languages including 13 Indian languages, and is as such able to handle linguistic inputs without the need to specify what the input language is upfront. The model produces similar embeddings as the bert-base-nli-stsb-mean-token model.

**Algorithm 1:** Semantic similarity between English and Malayalam

**Input:** A CSV file contains gloss and seed term of synset for languages English and Malayalam
**Output:** A CSV file contains id and semantic similarity score between English and Malayalam
**Data:** Common concepts in English and Malayalam languages
**foreach** *row in CSV file* **do**
  Compute embeddings of the English and Malayalam gloss
  Calculate cosine-similarity between the embeddings
  Write the id and semantic similarity score into a CSV file

**Algorithm 2:** Semantic similarity between Malayalam gloss and each word in the synset

**Input:** A CSV file contains gloss and a word for Malayalam language
**Output:** A CSV file contains id and semantic similarity score of Malayalam gloss and word
**Data:** Common concepts in English and Malayalam languages
**foreach** *row in CSV file* **do**
  Compute embeddings of the Malayalam gloss and word
  Calculate cosine-similarity between the embeddings
  Write the id and semantic similarity score into a CSV file

## 5 Results

There are 30139 Synsets available in the Malayalam language. Using our approach, we estimated that 14% of the Synsets have schema errors.

| | Gloss | | Word sense | |
|---|---|---|---|---|
| | **Eng-Mal** | **Hin-Mal** | **Eng-Mal** | **Hin-Mal** |
| Low(less than 0.4) | 30% | 4% | 17% | 2% |
| Moderate(0.4-0.7) | 55% | 35% | 59% | 28% |
| High(grater than 0.7) | 15% | 61% | 24% | 70% |

Figure 8: Summary of the similarity agreement of MWN between PWN and HWN

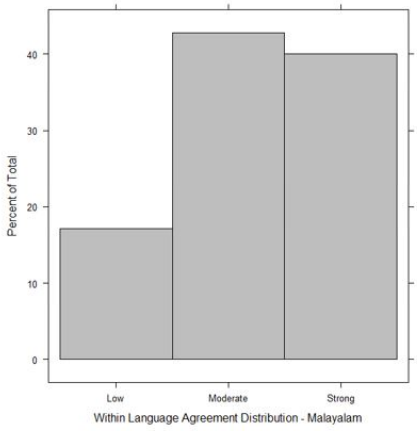



Figure 9: Distribution of similarity agreement within MWN

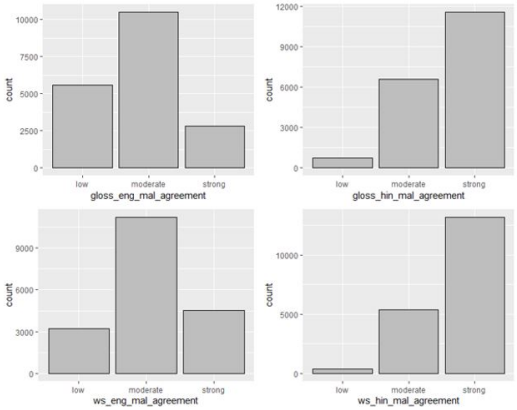



Figure 10: Distribution of similarity of MWN between PWN and HWN

Figure 8 summarizes cosine similarity score distribution. 30% of gloss-wise similarity scores for PWN-MWN synsets fall in the low alignment category, the score goes down to 17% if looking at word sense level alignment for PWN-MWN entries. Interestingly, when looking at HWN-MWN data, synsets falling into the low alignment category decrease to 4% when looking at gloss-wise alignment; 2% when looking at word sense-wise alignment. Looking at within language data, reflects that 20% of MWN word senses are tied to low scores(below 0.4). Figure 9 shows the distribution of similarity agreement within the Malayalam language and Figure 10 shows the similarity distribution of Malayalam language between English and Hindi.

Computing semantic similarity within the language helps us understand and isolate poorly aligned word senses and better word senses. Results aligned how Malayalam language entries are more semantically aligned with Hindi than English, which is expected.

## 6 Conclusion and Future work

This paper proposes specifications to be considered for a good quality lexical-semantic resource. We believe a resource without schema and semantic errors to be a useful resource. We defined schema errors and semantic errors for WordNet and estimated the presence of both for MWN. 14% of the Malayalam resource has schema errors. An average of 23.5% meanings representing the same concepts across English and Malayalam are tied to low similarity scores( poor semantic alignment). Only 3% of meanings representing the same concepts across Hindi and Malayalam have low similarity scores. Our approach concluded that when looking at within language data, 20% of MWN word senses are poorly aligned with their gloss. With this finding, we can classify MWN into two parts: Synsets requiring human revision and Synsets that can be readily shared with the world. We will be using MWN as part of a web application that supports end- Users in finding translations.

We are not taking these values as the final deciding factor. We will be using this low score synsets as a candidate set for our crowdsourcing application. This application will have different tasks like define gloss, provide Synset, validate the gloss, and so on.

## References


Ronald Asher. 2013. *Malayalam*. Routledge.

Ricardo Baeza-Yates, Luz Rello, and Julia Dembowski. 2015. CASSA: A context-aware synonym simplification algorithm. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1380–1385, Denver, Colorado. Association for Computational Linguistics.

Khuyagbaatar Batsuren, Gábor Bella, and Fausto Giunchiglia. 2019. Cognet: A large-scale cognate database. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3136–3145.

Gábor Bella, Linda Gremes, and Fausto Giunchiglia. 2020. Exploring the language of data. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6638–6648, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. 2014. Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, 56:85–100.

Nancy Bonvillain. 2019. *Language, culture, and communication: The meaning of messages*. Rowman & Littlefield.

Daren C Brabham. 2013. *Crowdsourcing*. Mit Press.

Bharathi Raja Chakravarthi. 2020. *Leveraging orthographic information to improve machine translation of under-resourced languages*. Ph.D. thesis, NUI Galway.

Bharathi Raja Chakravarthi, Mihael Arcan, and John P McCrae. 2018. Improving wordnets for under-resourced languages using machine translation. In *Proceedings of the 9th Global WordNet Conference (GWC 2018)*, page 78.

Bharathi Raja Chakravarthi, Mihael Arcan, and John P. McCrae. 2019. WordNet gloss translation for under-resourced languages using multilingual neural machine translation. In *Proceedings of the Second Workshop on Multilingualism at the Intersection of Knowledge Bases and Machine Translation*, pages 1–7, Dublin, Ireland. European Association for Machine Translation.

Bharathi Raja Chakravarthi, Navaneethan Rajasekaran, Mihael Arcan, Kevin McGuinness, Noel E. O'Connor, and John P. McCrae. 2020a. Bilingual lexicon induction across orthographically-distinct under-resourced Dravidian languages. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 57–69, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Bharathi Raja Chakravarthi, Priya Rani, Mihael Arcan, and John P McCrae. 2020b. A survey of orthographic information in machine translation. *arXiv preprint arXiv:2008.01391*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Niladri Sekhar Dash, Pushpak Bhattacharyya, and Jyoti D Pawar. 2017. *The WordNet in Indian Languages*. Springer.

Donald Foss. 1972. Psycholinguistics. *Psyccritiques*, 17(12).

Karimpumannil Mathai George. 1972. *Western influence on Malayalam language and literature*. Sahitya Akademi.

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*.

Daniel Gouadec. 2010. Quality in translation. *Handbook of translation studies*, 1:270–275.

Susan C Herring. 2008. Language and the internet. *The International Encyclopedia of Communication*.

Juliane House. 2014. Translation quality assessment: Past and present. In *Translation: A multidisciplinary approach*, pages 241–264. Springer.

Michael Ibba and Dieter Söll. 1999. Quality control mechanisms during translation. *Science*, 286(5446):1893–1897.

Mikhail Khodak, Andrej Risteski, Christiane Fellbaum, and Sanjeev Arora. 2017. Automated WordNet construction using word embeddings. In *Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications*, pages 12–23, Valencia, Spain. Association for Computational Linguistics.

Steven Krauwer. 2003. The basic language resource kit (blark) as the first milestone for the language resources roadmap. *Proceedings of SPECOM 2003*, pages 8–15.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Tradumàtica*, (12):0455–463.

Neil Mercer. 2002. *Words and minds: How we use language to think together*. Routledge.

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

Dan Moldovan and Vasile Rus. 2001. Logic form transformation of WordNet and its applicability to question answering. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 402–409, Toulouse, France. Association for Computational Linguistics.

Nandu Chandran Nair, Rajendran Sankara Velayuthan, and Khuyagbaatar Batsuren. 2019. Aligning the indowordnet with the princeton wordnet. In *Proceedings of the 3rd International Conference on Natural Language and Speech Processing*, pages 9–16.

Thomas Palakeel. 1996. Twentieth century malayalam literature. *Handbook of Twentieth-century Literatures of India*, pages 180–200.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, pages 8026–8037.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.

RI Prasanth. 2016. Lost word is lost world-a study of malayalam. *Language in India*, 16(6).

Faisal Rahutomo, Teruaki Kitasuka, and Masayoshi Aritsugi. 2012. Semantic cosine similarity. In *The 7th International Student Conference on Advanced Science and Technology ICAST*, volume 4.

S Rajendran and KP Soman. 2017. Malayalam wordnet. In *The WordNet in Indian Languages*, pages 119–145. Springer.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

T Santhosh, KG Varghese, R Sulochana, and R Kumar. 2002. Malayalam spell checker. In *Proceedings of the International Conference on Universal Knowledge and Language-2002*.

Sumaja Sasidharan, R Loganathan, and KP Soman. 2009. English to malayalam transliteration using sequence labeling approach. *International Journal of Recent Trends in Engineering*, 1(2):170.

S Sooraj, K Manjusha, M Anand Kumar, and KP Soman. 2018. Deep learning based spell checker for malayalam language. *Journal of Intelligent & Fuzzy Systems*, 34(3):1427–1434.

Ahmed Tawfik, Fausto Giunchiglia, and Vincenzo Maltese. 2014. A collaborative platform for multilingual ontology development. *World Academy of Science, Engineering and Technology*, 8(12):1.

Titela Vîlceanu. 2017. Quality assurance in translation. a process-oriented approach. *Romanian Journal of English Studies*, 14(1):141–146.

Mark Warschauer. 2002. Reconceptualizing the digital divide. *First monday*, 7(7).

Earnest Wheeler and Tawanna R Dillahunt. 2018. Navigating the job search as a low-resourced job seeker. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 48. ACM.

Zhi-Hua Zhou. 2016. Learnware: on the future of machine learning. *Frontiers Comput. Sci.*, 10(4):589–590.