Towards Producing Human-Validated Translation Resources for the Fula language through WordNet Linking

Khalil Mrini and Martin Benjamin
Ecole Polytechnique Fédérale de Lausanne
Switzerland
{khalil.mrini, martin.benjamin}@epfl.ch

Abstract

We propose methods to link automatically parsed linguistic data to the WordNet. We apply these methods on a trilingual dictionary in Fula, English and French. Dictionary entry parsing is used to collect the linguistic data. Then we connect it to the Open Multilingual WordNet (OMW) through two attempts, and use confidence scores to quantify accuracy. We obtained 11,000 entries in parsing and linked about 58% to the OMW on the first attempt, and an additional 14% in the second one. These links are due to be validated by Fula speakers before being added to the Kamusi Project's database.

1 Introduction

Multilingual dictionaries can be transformed to translation resources through Dictionary Entry Parsing (Lemnitzer and Kunze, 2005; Neff and Boguraev, 1989), that could be used for Machine Translation (Knight and Luk, 1994; Neff and McCord, 1990).

This paper describes first the conversion of a Fula[1] language dictionary (Fulfulde-English-French Lexicon, or FEFL) (Osborn et al., 1993), designed to be read as text, to a structured format that can be interoperable with other languages. The source is a trilingual lexicon offering translations to English and French for each entry. The resulting data is to be added to the Kamusi Project (Benjamin, 1995), which aims to collect linguistic data from many languages, with a special focus on African languages. The Fula language continuum is one of the major members of the Atlantic subfamily of the Niger-Congo languages (Ladefoged, 1968). Varieties, some of which are noted in the source dictionary, are spoken by 24 million people (Parkvall, 2007) in about 21 countries across Western and Central Africa.

To be able to connect Fula to other languages, it must be linked to a lexical base such as the Princeton WordNet (Fellbaum, 1998). Through this, the language is linked to the other languages available in the Open Multilingual WordNet (Bond and Paik, 2012; Bond and Foster, 2013).

This paper proposes a method to link entries collected from a multilingual lexicon to the WordNet. We evaluate each link using a confidence score giving an estimation of its ambiguity. The interoperability with other languages makes this lexicon a language resource that translators and interpreters can use. Finally, this work aims to prepare the collected data for future validation by humans.

2 Parsing

The parsing of a dictionary requires first an analysis of its format. Moreover, both format and content need to be made compatible with Kamusi's own and the data needs to be filtered for relevant categories. In this case, the authors could read the English and French elements of the source dictionary, but had no familiarity with the Fula language. Nor does another data source exist that could shed light on the Fula content due to rare electronic resources. Fortunately, FEFL's lead author provided the lexicon in a machine-readable format.

In this section, we first describe Kamusi's work history and why this resource is emblematic for languishing linguistic data. Then we elaborate on the source dictionary and the parsing method used.

[1] Also known as Fulah, Fulani, Fulfulde, Peul, Pulaar, and Pular. The macrolanguage has ISO 639-3 designation "*ful*", with nine variations assigned individual codes by Ethnologue. https://www.ethnologue.com/subgroups/fula-1

The Proceedings of the First Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT), pages 58–64,
Varna, Bulgaria, Sept 7, 2017
https://doi.org/10.26615/978-954-452-042-7_008

```
ABADA Ar
abada, abadaa, abadan DFZ Z<->
never(F) (with negation); ever(F); long ago
jamais(D) (avec la négation)(Z); jamais; il y a longtemps
Abada mi yahaali.  (F): I have never gone.  ; Je ne suis jamais allé.
abada pati (F): don't ever ; ne faîtes jamais (qqch)
gila abada (F): since long ago, forever ; depuis longtemps, toujours
```

Figure 1: Example of an entry in the Fula dictionary

2.1 Kamusi

The goal of Kamusi is to assemble as much linguistic data as possible in a highly structured format under a single umbrella that can be accessed by the public and as linked data for Natural Language Processing (NLP). Within each language, individual senses of each term are considered their own entities. Terms are then joined at the semantic level across languages (with attention to semantic drift and lexical gaps).

The project started with Swahili, and the multilingual expansion was originally planned with a focus on other African languages. As the model was developed and data collection started, though, African languages got pushed toward the rear because no data was available in digital form, or because these languages might have at best a bespoke bilingual electronic or print dictionary with English or French. This resource is therefore a way for Kamusi to strengthen its focus on African languages and address the scarcity of digitally ready African linguistic data.

Even after getting the data and overcoming the challenges for parsing and aligning data, it remains difficult to perform word sense disambiguation automatically (Ide and Véronis, 1998; Lesk, 1986; Navigli, 2009; Rigau and Agirre, 1995). Disambiguation requires human attention, for which the DUCKS (Data Unified Conceptual Knowledge Sets) tool has been developed and is being tested, but it needs resources to develop groups that can work with the lexicons of their languages.

2.2 The Source Dictionary

The source dictionary was begun in 1989. The FEFL authors transmitted the dictionary document for incorporation within Kamusi without copyright restriction. For parsing, the document was converted to plain text.

The FEFL is ordered by the Fula root, with separate entries for each derivative. As a text document, this was a logical way of structuring related Fula terms. However, within our data structure each sense of each word is its own entity, with a feature like "*root*" as one element of the data. Finding all descendants of a common root becomes a function of the search query, rather than a guiding organizational principle.

Each FEFL entry contains at least three lines: first Fula, then English, and finally French. Sometimes, an entry can simply be a cross-reference to the root, performed in one line. That entry might also have lemmas that could be useful for collection. Importantly, as with many multilingual dictionaries, the entries do not contain own-language definitions, but rather ascribe meaning in relation to the given English and French equivalents, and oftentimes Fula usage examples and their translations.

The Fula line begins with at least one Fula lemma and information on the sources, using abbreviations and whether the source ascribes the word to one or more dialects. The Fula language is a continuum with questionable inter-intelligibility from its eastern to western extremes, and it is important to retain the information on dialects as the base for future research. The Fula line also gives abbreviated information on the part-of-speech (PoS) tag. An annex to the dictionary explains all the abbreviations.

The Fula line is followed by the English and then French line, also separated by commas or semicolons. These lines may optionally be followed by annotation lines.

The line for the roots is easily recognizable because the root is written in block capital letters. However, sometimes the line may indicate suffixes to the previous root or a new root. It may also include information on the etymologic origin of the word.

Taking into account the dictionary's specificities is necessary to automatically parse all the en-

tries. An example of an entry is in Figure 1. This example has lemmas in Fula (second line), English (third line) and French (fourth line) with information on sources in parentheses, a line for the root including dialect information (first line) and three lines of annotations at the end.

2.3 Parsing Method

We parsed the source dictionary with a method that evolved as we were able to make sense of the data. It evaluates each non-empty line. We first initiate a new Fula entry. If the current line is not referencing another entry, then there are two cases.

The first case is when the line is a root line. If the Fula entry is complete, meaning it has a root, a Fula line, an English line and a French line, the filtered data is printed into tab-separated text files and a new Fula entry is set. If the line starts with a dash and the current entry's root is non-empty, the suffix is added to that root. Otherwise, the line contains a new root.

The second case is when none of the conditions for the last two have been fulfilled. Then there are two subcases.

The first subcase is when the Fula entry is complete with a root, Fula, English and French lines, then a check is run on the line to see if it is an annotation line that has to be added to the current entry. If the line is instead a line containing at the same time a root and a word, it is ignored. Otherwise, it must be the Fula line of the next entry. Afterwards, the filtered data is first saved and a new Fula entry is initiated with the same root as the previous one and the Fula line is added to it.

The second subcase is when the Fula entry is not complete. If the current line is not an annotation line, it contains either the English or French line, and it is added to the current entry. If the line is found to be an annotation line, the entry is deficient and therefore has to be deleted. We then start looking for a new Fula entry, and this new entry's root is the same as the previous one, unless the next line is a root line.

These two cases ensure all valid Fula entries are collected. However, when valid lines are collected, they are transformed to be cleaned of unnecessary information and separated from information that is considered useful to the preponderance of online dictionary users. The relevant information that is kept is dialects, synonyms that are the lemmas shown in brackets, and PoS tags.

Inside the English and French lines, rough synonyms are separated by commas while different senses are separated by semicolons. The English and French lines both have the same number of synonym sets in the same order, though not necessarily the same number of terms for each concept. The program can thus separate senses into different entries on the base of semicolons, but cannot definitively match specific English terms to specific French terms within synonym sets that can be recognized to share a general topical meaning. For each sense, English information in parentheses is preserved.

At the end, each Fula entry has an ID and inside each entry, each sense has an ID. Eleven tab-separated text files are printed: one for annotations, one for dialects, one for entries that display the Fula line followed by the English and French lines, one for Fula lemmas, one for PoS tags, one for roots, one for sense annotations, one for sense classifications, one for the English sense, one for the French sense and finally one for Fula synonyms. When parsing was completed, the source dictionary resolved to 7918 Fula entries and 10970 Fula senses.

3 Linking to the WordNet

A main purpose of bringing the FEFL data into Kamusi is to make it interoperable with other languages that exploit the same technology. In the case of Fula, this will result in translation resources with neighboring languages such as Songhay and Bambara that have not heretofore been possible. To achieve these objectives, the Fula terms must be connected to the overarching concept sets that Kamusi uses to establish semantic links across languages. Kamusi uses the roughly 100,000 synset definitions from the Princeton WordNet as the starting point for aligning concepts. The nearly 11,000 Fula senses obtained through the parsing procedures described in the previous section can join a larger multilingual database, that is the Open Multilingual WordNet, by being linked to the Princeton WordNet.

3.1 The Princeton WordNet and the Open Multilingual WordNet

The Princeton WordNet (PWN) is an electronic lexical database created in the Cognitive Science Laboratory of Princeton University (Fellbaum, 1998) that separates terms by their senses, and

joins terms in "*synsets*" (unordered sets of rough synonyms) that share a general definition. The WordNet concept has now been applied to many other languages, using the PWN synsets as the base set of concepts to populate with the lemmas for their own equivalent terms.

The Open Multilingual WordNet (OMW) (Bond and Foster, 2013) is a collection of standalone wordnets from several dozen languages that could have teams to produce them and have chosen to share their work, with most of their synsets indexed to PWN. If terms from any non-wordnet language can be matched to defined concepts in PWN, they can thus be joined as rough bilingual matches across the OMW.

WordNet divides synsets into four main categories: nouns, verbs, adjectives and adverbs. However, it does not reference function words like prepositions and determiners. So the earlier four parts of speech are the ones that were used to link PWN synsets and the FEFL senses.

3.2 Related Work

Most freely available wordnets use the *expand* method (Vossen, 2005) by adding new lemmas to the existing synsets in the Princeton WordNet. Although Fellbaum and Vossen (2012) argue that this is an imperfect method that poses the question of equivalence, it is useful for this case because FEFL is intended to be understood in reference to the stock of English translation equivalents.

Other wordnets have used the *merge* approach, which Balkova et al. (2004) define as "*building taxonomies from monolingual lexical resources and then, making a mapping process using bilingual dictionaries*". It was used by wordnets such as the Urdu one (Zafar et al., 2012, 2014), whereas the EuroWordNet (Vossen, 1998) is an example of a wordnet using a mixture of both methods. The EuroWordNet also proposed an interlingual index (ILI) (Fellbaum and Vossen, 2008) to tackle concept equivalence between the different languages it contains, whereas Bond et al. (2016) propose a collaborative form of the ILI (CILI) to extend it to all other languages. Kotis et al. (2006) propose an automatic merge approach making use of Latent Semantic Indexing (LSI) (Hofmann, 1999).

3.3 WordNet-linking Method

To understand the method easily, we provide the flowchart in Figure 2 as illustration. Examples of Fula entries will also be used. The Fula word "*adadu*" has the English definition "*quantity, measure; sum, total; calculation; number*". One can notice that senses were separated by a semicolon and synonym terms of the same sense are separated by a comma. In this method, there were two attempts to connect the Fula data through the English translations to the WordNet. The first one considered the senses as separated by semi-colon (step **a** in Figure 2). The second one was more flexible and considered separating senses even further by commas (step **i**). The confidence score formula was adapted to penalize flexibility, as it diminishes accuracy.

In both attempts, the PoS tags were used to identify ID lists of verbs, adjectives and adverbs. Given that 70.4% of senses in the WordNet are nouns, it made sense to have it as the default PoS tag. These lists were used to search for corresponding synsets with the matching PoS tag. For each definition, words were tokenised and stop words were removed unless there was only one word in the definition.

In the first attempt, senses were separated by semicolons. In the above example, 4 senses were obtained. Then, in each sense, the words that were not separated by a comma were joined by an underscore to search for a multiple-word expression in the WordNet (step **b**). For example, the Fula word "*aadamanke*" has the English definition "*human being*". The WordNet was queried for "*human_being*" and gave a set of one synset. However, if this query gave an empty set, then individual words "*human*" and "*being*" would have been matched to the WordNet as in step **c** and a set of synsets is given for each word. Only synsets present in all of the sets were kept (step **d**). This intersection of all non-empty sets became the set of synsets for that sense.

In the instance of the Fula verb "*aatude*", the English definition "*scream loudly, cry out*" has two parts. The first part "*scream loudly*" matches to 3 synsets (step **c**). The second part "*cry out*" matches to 7 synsets (step **b**). They overlap in 1 synset, which will therefore be the only one matching the whole definition (step **d**). Since this final result is determined by more than one non-empty set of synsets, then it is considered the result of an intersection (steps **f** and **h**).

If the final set is an intersection of sets of multiple sub-senses, then there is more confidence in the WordNet matches obtained and so we decided

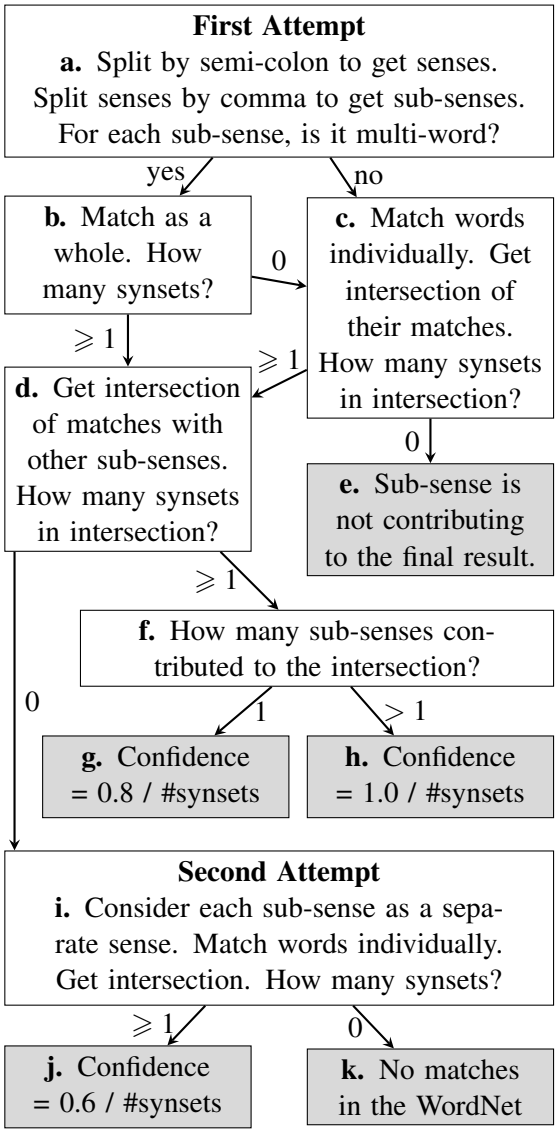

Figure 2: Flowchart of the WordNet-linking method, with final states in gray

to set the confidence score at 1.0 divided by the number of intersecting synsets. If the final set is not an intersection, and therefore was the result of at most a few words not separated by commas, the inaccuracy of not being confirmed multiple times must be penalized. So the confidence score given is 0.8 divided by the number of synsets in the final set (step **g**).

In the second attempt (step **i**), the Fula entries considered are those left unconnected to the WordNet in the prior attempt. The words in the senses, here defined by separation both by commas and semicolons, are not joined by an underscore for WordNet matching but are rather matched individually. So the final set will be the intersection of these words' synset matches. For instance, the Fula verb "*aalude*" has English definition "*split, dissociate*". This sense has two parts when we split by comma. No common synset could be found between the two parts. Therefore, we split the sense by comma and obtained two senses "*split*" and "*dissociate*". These have separate matches in the WordNet and therefore the confidence score is also separate.

This second attempt is more flexible than the first one. So for each sense it matched, the confidence score will be 0.6 divided by the number of synsets in the final set (step **j**). The confidence scores were computed such that the greater the ambiguity, the lower the score. Items that have only one match to the WordNet can be clearly distinguished, as their scores will be either 1, 0.8 or 0.6. Meanwhile, items that have multiple WordNet matches (0.5 or below) are quickly filtered out to diminish ambiguity. In the end, the confidence scores proved useful in determining whether an entry could be accepted as-is, or placed in Kamusi's DUCKS tool for human review.

3.4 Results and Discussion

The links automatically established by the WordNet-linking method are in Table 1. 72.4% of all Fula senses were linked to the WordNet. Links with confidence score 1.0 indicate an almost-certain match, whereas links with confidence score 0.8 or 0.6 indicate likely matches. At the end, 3031 Fula senses (27.6% of total) remained without any potential WordNet connections.

Such examples of Fula words that ended up without WordNet connections include pronouns (such as "*you*") that the WordNet does not include. Because some non-noun words were not PoS-tagged in the FEFL and because of the assumption that all entries without PoS tags were nouns, non-PoS-tagged entries such as "*never*" and "*ever*" that are adverbs could not be matched. In other cases, matches were not found between concepts because the sources use different terms to render a similar idea, such as "*person who is knowledgeable*" in the FEFL versus "*wise man*" in PWN. Still other non-matches are due to different patterns for expressing concepts that have a shared cultural existence, such as the verb "*seyadde*", that in English is "*be*" plus the adjective "*happy*".

However, a large (uncounted) number of unmatched Fula words are very specific to the Central and Western African context. Such words are

Attempt	Initial senses	Senses linked	**Confidence score**: Senses with 1 link
First	10970	58.3% (6391)	**1.0:** 5.2% (332); **0.8:** 20.3% (1295)
Second	4579	3543 sub-definitions linked, which resulted from 1548 senses (14.1%)	**0.6:** 16.4% (581)

Table 1: Results of the two WordNet-linking attempts as applied on the FEFL senses

Figure 3: Live version of DUCKS, with the Comparative African Word List (CAWL) as active dataset

for instance the verb "*furraade*" which has the English translation "*break the fast at sunset*", or the noun "*maari*" which means in English "*condiment made from seeds of the locust bean tree*". From the perspective of Digital Humanities, these words that do not have a linguistic or conceptual equivalent in English are perhaps the most interesting result of mining a dictionary that grew from field lexicography, revealing indigenous concepts and making them visible to the global knowledge base.

3.5 Future Work

Subsequent steps include: making the data searchable online with its original trilingual sets and validating the data by humans through DUCKS.

DUCKS has been developed so that players are presented with a term in their native language on one side of their screen, and a list of WordNet senses for the given English equivalent. Then, players can chose which senses match the term, as in Figure 3. This crowd-sourced validation can replace the one performed by authors of wordnets such as the Japanese (Isahara et al., 2008) and Arabic (Black et al., 2006) ones. The success rate of our algorithm will be determined by the number of WordNet links approved by Fula speakers.

The senses with no match to the WordNet are ineligible for DUCKS until further human review, that might establish other existing English terms for alignment.

4 Conclusions

This paper proposed methods to collect linguistic data automatically using dictionary entry parsing and wordnet linking. We applied these methods to a trilingual Fula-English-French lexicon (FEFL) (Osborn et al., 1993).

First, a thorough analysis of the format of the dictionary was necessary in order to parse it and collect the necessary data, with the method being refined empirically. At the end, the parsing resulted in 7918 Fula entries and 10970 Fula senses gathered, organised in 11 categories of useful data.

Then, to provide a base for semantic comparison, the Fula data was linked to the Princeton WordNet (Fellbaum, 1998). Through this linking, it is connected to all languages available in the Open Multilingual WordNet (Bond and Foster, 2013). Two attempts were made, with the second one being more flexible. Confidence scores were given to each match, to gauge their accuracy. The first attempt scored 6391 potential matches whereas the second one scored 3543 matches. In total, 72.4% of the Fula senses were linked. Many of the 3031 unmatched Fula senses were related to the specific cultural and geographical context where the language is used.

This automatically collected and linked translation resource will be put in DUCKS to be validated by Fula speakers, before joining Kamusi data.

References

Valentina Balkova, Andrey Sukhonogov, and Sergey Yablonsky. 2004. Russian wordnet. In *Proceedings of the Second Global Wordnet Conference*.

Martin Benjamin. 1995. Kamusi gold (global online living dictionary). Accessed: 2017-08-04. https://kamusi.org/.

William Black, Sabri Elkateb, Horacio Rodriguez, Musa Alkhalifa, Piek Vossen, Adam Pease, and Christiane Fellbaum. 2006. Introducing the arabic wordnet project. In *Proceedings of the third international WordNet conference*. pages 295–300.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *ACL (1)*. pages 1352–1362.

Francis Bond and Kyonghee Paik. 2012. A survey of wordnets and their licenses. *Small* 8(4):5.

Francis Bond, Piek Vossen, John P McCrae, and Christiane Fellbaum. 2016. Cili: the collaborative interlingual index. In *Proceedings of the Global WordNet Conference*. volume 2016.

Christiane Fellbaum. 1998. *WordNet*. Wiley Online Library.

Christiane Fellbaum and Piek Vossen. 2008. Challenges for a global wordnet. In *Online Proceedings of the First International Workshop on Global Interoperability for Language Resources*. pages 75–82.

Christiane Fellbaum and Piek Vossen. 2012. Challenges for a multilingual wordnet. *Language Resources and Evaluation* 46(2):313–326.

Thomas Hofmann. 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, pages 50–57.

Nancy Ide and Jean Véronis. 1998. Introduction to the special issue on word sense disambiguation: the state of the art. *Computational linguistics* 24(1):2–40.

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

Kevin Knight and Steve K Luk. 1994. Building a large-scale knowledge base for machine translation. In *AAAI*. volume 94, pages 773–778.

Konstantinos Kotis, George A Vouros, and Konstantinos Stergiou. 2006. Towards automatic merging of domain ontologies: The hcone-merge approach. *Web semantics: Science, services and agents on the world wide web* 4(1):60–79.

Peter Ladefoged. 1968. *A phonetic study of West African languages: An auditory-instrumental survey*. 1. Cambridge University Press.

Lothar Lemnitzer and Claudia Kunze. 2005. Dictionary entry parsing. *ESSLLI-2005* .

Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*. ACM, pages 24–26.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)* 41(2):10.

Mary S Neff and Branimir K Boguraev. 1989. Dictionaries, dictionary grammars and dictionary entry parsing. In *Proceedings of the 27th annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, pages 91–101.

Mary S Neff and Michael C McCord. 1990. *Acquiring lexical data from machine-readable dictionary resources for machine translation*. IBM Thomas J. Watson Research Division.

Donald W. Osborn, David J. Dwyer, and Joseph I. Donohoe Jr. 1993. *A Fulfulde (Maasina)-English-French Lexicon: A Root-based Compilation Drawn from Extant Sources Followed by English-Fulfulde and French-Fulfulde Listings*. Michigan State University Press.

Mikael Parkvall. 2007. Världens 100 största språk 2007. *The World's* 100.

German Rigau and Eneko Agirre. 1995. Disambiguating bilingual nominal entries against wordnet. *arXiv preprint cmp-lg/9510004* .

Piek Vossen. 1998. Introduction to eurowordnet. *Computers and the Humanities* 32(2-3):73–89.

Piek Vossen. 2005. Building wordnets. Accessed: 2017-08-07. http://www.globalwordnet.org/gwa/BuildingWordnets.ppt.

Ayesha Zafar, Afia Mahmood, Farhat Abdullah, Saira Zahid, Sarmad Hussain, and Asad Mustafa. 2012. Developing urdu wordnet using the merge approach. In *Proceedings of the Conference on Language and Technology*. pages 55–59.

Ayesha Zafar, Afia Mahmood, Sana Shams, and Sarmad Hussain. 2014. Structural analysis of linking urdu wordnet to pwn 2.1. In *the Proceedings of Conference on Language and Technology 2014 (CLT14)*.