# Semi-Automatic Development of KurdNet, The Kurdish WordNet

**Purya Aliabadi**
SRBIAU
Sanandaj, Iran
puryait@gmail.com

## Abstract

Recently, we reported on our efforts to build the first prototype of KurdNet. In this proposal, we highlight the shortcomings of the current prototype and put forward a detailed plan to transform this prototype to a full-fledged lexical database for the Kurdish language.

## 1 Introduction

WordNet (Fellbaum, 2010) has been used in numerous natural language processing tasks such as word sense disambiguation and information extraction with considerable success. Motivated by this success, many projects have been undertaken to build similar lexical databases for other languages. Among the large-scale projects are EuroWordNet (Vossen, 1998) and BalkaNet (Tufis et al., 2004) for European languages and IndoWordNet (Bhattacharyya, 2010) for Indian languages.

Kurdish belongs to the Indo-European family of languages and is spoken in Kurdistan, a large geographical region spanning the intersections of Iran, Iraq, Turkey, and Syria (as showed in Figure 1). Kurdish is a less-resourced language for which, among other resources, no wordnet has been built yet.

Despite having a large number (20 to 30 millions) of native speakers (Hassanpour et al., 2012; Haig and Matras, 2002), Kurdish is among the less-resourced languages for which the only linguistic resource available on the Web is raw text (Walther and Sagot, 2010). In order to address this resource-scarceness problem, the Kurdish language processing project (KLPP[1]) has been recently launched at University of Kurdistan. Among the the major linguistic resources that KLPP has been trying to develop is KurdNet, a

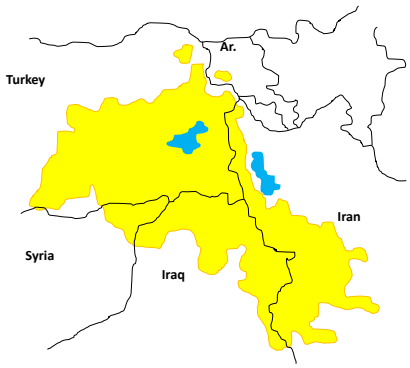


Figure 1: Geographical Distribution of Kurdish Speakers

WordNet-like lexical database for the Kurdish language. Earlier this year, we reported (Aliabadi et al., 2014) on our effort to build the first prototype of KurdNet. In this paper, we propose a plan to transform this preliminary version into a full-fledged and functional lexical database.

The rest of this paper is organized as follows. We first (in Section 2) give a brief overview of the current state of KurdNet. Then after highlighting the main shortcomings of the current prototype in Section 3, we present our plan to transform this prototype to a full-blown lexical database for the Kurdish language in Section 4. We conclude the paper in Section 5.

## 2 KurdNet: State-of-the-Art

In our previous work (Aliabadi et al., 2014), we described the steps that we have taken to build the first prototype of KurdNet. There, we

1. highlighted the main challenges in building a wordnet for the Kurdish language (including its inherent diversity and morphological complexity),

2. built the first prototype of KurdNet, the Kurdish WordNet (see a summary below), and

[1] http://eng.uok.ac.ir/esmaili/research/klpp/en/main.htm



*Proceedings of the ACL 2014 Student Research Workshop*, pages 94–99,
Baltimore, Maryland USA, June 22-27 2014. 

3. conducted a set of experiments to evaluate the impact of KurdNet on Kurdish information retrieval.

In the following, we first define the scope of our first prototype, then after justifying our choice of construction model, we describe KurdNet's individual elements.

### 2.1 Scope

Kurdish has two main dialects (Esmaili and Salavati, 2013): Sorani and Kurmanji. In the first prototype of KurdNet we focus only on the Sorani dialect. This is mainly due to lack of an available and reliable Kurmanji-to-English dictionary. Moreover, processing Sorani is in general more challenging than Kurmanji (Esmaili et al., 2013a).

### 2.2 Methodology

There are two well-known models for building wordnets for a language (Vossen, 1998):

- **Expand**: in this model, the synsets are built in correspondence with the WordNet synsets and the semantic relations are directly imported. It has been used for Italian in MultiWordNet and for Spanish in EuroWordNet.
- **Merge**: in this model, the synsets and relations are first built independently and then they are aligned with WordNet's. It has been the dominant model in building BalkaNet and EuroWordNet.

The expand model seems less complex and guarantees the highest degree of compatibility across different wordnets. But it also has potential drawbacks. The most serious risk is that of forcing an excessive dependency on the lexical and conceptual structure of one of the languages involved, as pointed out in (Vossen, 1996).

In our project, we follow the Expand model, since it can be partly automated and therefore would be faster. More precisely, we aim at creating a Kurdish translation/alignment for the Base Concepts (Vossen et al., 1998) which is a set of 5,000 essential concepts (i.e. synsets) that play a major role in the wordnets. Base Concepts (BC) is available on the Global WordNet Association (GWA)'s Web page[2]. The Entity-Relationship (ER) model for the data represented in Base Concept is shown in Figure 2. A sample synset is depicted in Figure 3.

[2]http://globalwordnet.org/

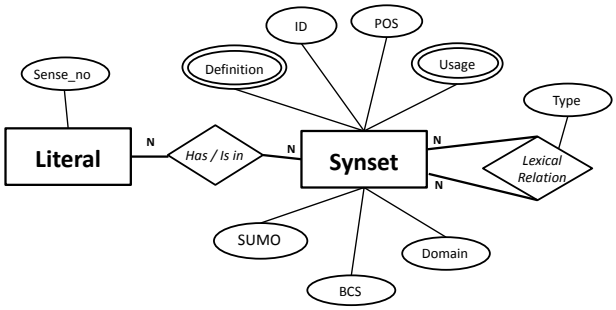


Figure 2: Base Concepts' ER Model (Aliabadi et al., 2014)

```
<SYNSET>
    <ID>ENG20-00008853-v</ID>
    <POS>v</POS>
    <SYNONYM>
        <LITERAL>shed<SENSE>4</SENSE></LITERAL>
        <LITERAL>molt<SENSE>1</SENSE></LITERAL>
                <LITERAL>exuviate<SENSE>1</SENSE></LITERAL>
        <LITERAL>moult<SENSE>1</SENSE></LITERAL>
        <LITERAL>slough<SENSE>1</SENSE></LITERAL>
        </SYNONYM>
    <ILR><TYPE>hypernym</TYPE>ENG20-01471089-v</ILR>
    <ILR><TYPE>eng_derivative</TYPE>ENG20-01245451-n</ILR>
    <ILR><TYPE>eng_derivative</TYPE>ENG20-08844332-n</ILR>
    <ILR><TYPE>eng_derivative</TYPE>ENG20-12753095-n</ILR>
    <ILR><TYPE>eng_derivative</TYPE>ENG20-12791455-n</ILR>
    <DEF>cast off hair, skin, horn, or feathers</DEF>
    <USAGE>out dog sheds every Spring</USAGE>
    <BCS>2</BCS>
        <DOMAIN>zoology</DOMAIN>
    <SUMO>Removing<TYPE>+</TYPE></SUMO>
</SYNSET>
```

Figure 3: A WordNet verb synset in XML (Vossen et al., 1998)

### 2.3 Elements

Since KurdNet follows the Expand model, it inherits most of Base Concepts' structural properties, including: synsets and the lexical relations among them, POS, Domain, BCS, and SUMO. KurdNet's language-specific aspects, on the other hand, have been built using a semi-automatic approach. Below, we elaborate on the details of construction the remaining three elements.

**Synset Alignments**: for each synset in BC, its counterpart in KurdNet is defined semi-automatically. We first use Dictio (a Sorani-English dictionary, see Section 4.2) to translate its literals (words). Having compiled the translation lists, we combine them in two different ways: (i) a maximal alignment (abbr. **max**) which is a *superset* of all lists, and (ii) a minimal alignment (abbr. **min**) which is a *subset* of non-empty lists. Figure 4 shows an illustration of these two combination variants. In future, we plan to apply

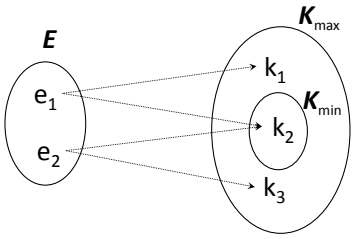


Figure 4: An Illustration of a Synset in Base Concepts and its Maximal and Minimal Alignment Variants in KurdNet (Aliabadi et al., 2014)

| | **Base Concepts** | **KurdNet** (max) | **KurdNet** (min) |
|---|---|---|---|
| Synset No. | *4,689* | *3,801* | *2,145* |
| Literal No. | *11,171* | *17,990* | *6,248* |
| Usage No. | *2,645* | *89,950* | *31,240* |

Table 1: The Main Statistical Properties of Base Concepts and its Alignment in KurdNet (Aliabadi et al., 2014)

more advanced techniques, similar to the graph algorithms described in (Flati and Navigli, 2012).

**Usage Examples**: we have taken a corpus-assisted approach to speed-up the process of providing usage examples for each aligned synset. To this end, we: (i) extract all sentences (820,203) of the Pewan corpus (Esmaili and Salavati, 2013), (ii) lemmatize the corpus to extract all the lemmas (278,873), and (iii) construct a lemma-to-sentence inverted index. In the current version of KurdNet, for each synset we build a pool of sentences by fetching the first 5 sentences of each of its literals from the inverted list. These pools will later be assessed by lexicographers to filter out non-relevant instances. In future, more sophisticated approaches can be applied (e.g., exploiting contextual information).

**Definitions**: due to lack of proper translation tools, this element was aligned manually. We built a graphical user interface to facilitate the lexicographers' task.

Table 1 shows a summary of KurdNet's statistical properties along with those of Base Concepts.

The latest snapshot of KurdNet's prototype is freely accessible and can be obtained from (KLPP, 2013).

| **Noun** | **Verb** | **Adjective** | **Adverb** |
|---|---|---|---|
| *Antonym* | *Antonym* | *Antonym* | *Antonym* |
| *Hyponym* | *Troponym* | *Similar* | *Derived* |
| *Hypernym* | *Hypernym* | *Relational Adj* | |
| *Meronym* | *Entailment* | *Also See* | |
| *Holonym* | *Cause* | *Attribute* | |

Table 2: WordNet Relational (Beckwith et al., 1993)

## 3 KurdNet: Shortcomings

The current version of KurdNet is quite basic and therefore its applicability is very limited. In order to expand the usability of KurdNet, the following shortcomings must be overcome:

### 3.1 Incomplete Coverage of Kurdish Vocabulary

KurdNet has been built as an alignment for Base Concepts and since Base Concepts contains only a small subset of English vocabulary, KurdNet's coverage is inevitably small. Furthermore, as it can be seen in Table 1, due to the limitations of the dictionaries used, not all English words in the Base Concepts (Vossen et al., 1998) have an equivalent in KurdNet. Hence the current mapping between WordNet and KurdNet is only partial. Finally, the lexical idiosyncrasies between Kurdish and English should be identified and included in KurdNet.

### 3.2 Refinement of Automatically-Generated Content

Each synset must contain a comprehensive definition and a practical example. While KurdNet definitions are provided manually and therefore enjoy high quality, the actual words in each synset as well as the usage examples have been produced manually. In order to increase the reliability and correctness of KurdNets, there need to be mechanisms to refine the existing machine-generated components.

### 3.3 Limited Support for Semantic Relation Types

As shown in Table 2, there are several WordNet semantic relations for each syntactic categories. Each syntactic categories are organized to component files (Miller et al., 1993). The most important semantic relation in WordNet is Hyponymy and this relation is the only one support in KurdNet (Aliabadi et al., 2014).

### 3.4 Absence of Kurmanji Synsets

Kurdish is considered a *bi-standard*[3] language (Gautier, 1998; Hassanpour et al., 2012): the **Sorani** dialect written in an Arabic-based alphabet and the **Kurmanji** dialect written in a Latin-based alphabet. The linguistics features distinguishing these two dialects are phonological, lexical, and morphological. The important morphological differences that concern the construction of KurdNet are (MacKenzie, 1961; Haig and Matras, 2002): (i) in contrast to Sorani, Kurmanji has retained both gender (feminine v. masculine) and case opposition (absolute v. oblique) for nouns and pronouns, and (ii) while is Kurmanji passive voice is constructed using the helper verb "hatin", in Sorani it is created via verb morphology. As explained in Section 2, the current KurdNet prototype only covers the Sorani dialect and therefore it should be extended to include the Kurmanji dialect as well. This would require not only using similar resources to those reported in this paper, but also building a mapping system between the Sorani and Kurmanji dialects.

### 3.5 Dictionary Imperfections

Dictio, the dictionary that was used for building KurdNet, is relatively small. We have recently discovered new linguistics resources that can improve the quality of automatic translation of English words and sentences into Kurdish and vice versa (see Section 4.2).

## 4 KurdNet: Extension Plan

### 4.1 Goals and Envisioned Outcomes

The main objectives and expected artefacts for this proposals are the following:

- to refine the current prototype, through use of intelligent algorithms and/or manual assistance.
- to winden the scope (i.e., including Kurmanji synsets), the coverage (i.e., going beyond Base Concepts) , and richness (supporting additional semantic relations) of the current version.
- to produce tool kits for users (e.g. graphical interfaces), developers (e.g., drivers and programming interfaces), and contributors (e.g., navigation/edition tools).
- to design and conduct experiments in order to assess the effectiveness of KurdNet in NLP and IR applications.
- to publish the innovative aspects as research papers.

### 4.2 Available Resources

Below are the Kurdish language resources that can be potentially used throughout this project:

- **KLPP Resources**
  – *the Pewan corpus* (Esmaili and Salavati, 2013): for both Sorani and Kurmanji dialects. Its basic statistics are shown in Table 3
  – *the Renoos lemmatizer* (Salavati et al., 2013): it is the result of a major revision of Jedar, a Kurdish stemmer whose outputs are stems.
  – *the Pewan test collection* (Esmaili et al., 2013b): is a test collection for both Sorani and Kurmanji.
- **Online Dictionaries:**
  – *Dictio*: an English-to-Sorani dictionary with more than 13,000 headwords. It employs a collaborative mechanism for enrichment.
  – *Ferheng*: a collection of dictionaries for the Kurmanji dialect with sizes ranging from medium (around 25,000 entries, for German and Turkish) to small (around 4,500, for English).
  – *Inkurdish*[4]: a new and high-quality translation between Sorani Kurdish and English.
  – *English Kurdish Translation*[5]: especially can translate words in Kurmanji and English together.
  – *Freelang*[6]:supports 4000 words in kurmanji.
  – *Glosbe*[7]: is a multilingual dictionary, that includes Soranii, Kurmanj, and English .
  – *Globalglossary*[8] is a Kurdish-English dictionary.

[3] Within KLPP, our focus has been on Sorani and Kurmanji which are the two most widely-spoken and closely-related dialects (Haig and Matras, 2002; Walther and Sagot, 2010).

[4] http://www.inkurdish.com
[5] http://www.englishkurdishtranslation.com/
[6] http://www.freelang.net/online/kurdish.php
[7] http://glosbe.com/en/ku/
[8] http://www.globalglossary.org/en/en/kmr/

| | Sorani | Kurmanji |
|---|---|---|
| Articles No. | *115,340* | *25,572* |
| Words No. (dist.) | *501,054* | *127,272* |
| Words No. (all) | *18,110,723* | *4,120,027* |

Table 3: The Pewan Corpus' Basic Statistics (Esmaili and Salavati, 2013)

- **Wikipedia**
  It currently has more than 12,000 Sorani[9] and 20,000 Kurmanji[10] articles. One useful application of these entries is to build a parallel collection of named entities across both dialects.

### 4.3 Methodology

As mentioned in Section 2, we have adopted the Expand model to build KurdNet. According to (Vossen, 1996), the MultiWordNet (MWN[11]) model (Expand model) seems less complex and guarantees the highest degree of compatibility across different wordnets. The MWN model also has potential drawbacks. The most serious risk is that of forcing an excessive dependency on the lexical and conceptual structure of one of the languages involved, as (Vossen, 1996) points out. This risk can be considerably reduced by allowing the new wordnet to diverge, when necessary, from the PWN.

Another important advantage of the MWN model is that automatic procedures can be devised to speed up both the construction of corresponding synsets and the detection of divergences between PWN and the wordnet being built. According to the Expand model, the aim is to build, whenever possible, Kurdish synsets which are synonymous (semantically correspondent) with the PWN synsets. The second strategy is based on Kurdish-to-English translations. For each sense of a Kurdish word K, we look for a PWN synset S including at least one English translation of K and a link between K and S is established (Pianta et al., 2002).

For the correct alignment of Sorani and Kurmanji synsets, we propose to use three complementary approaches:

- use of English (here, Base Concepts) synsets as reference points between both dictionary-translated synsets of Sorani and Kurmanji.

| English | Sorani | Kurmanji |
|---|---|---|
| *word1* | *S-translation1* | *K-translation1* |
| *word2* | *S-translation2* | *K-translation2* |
| *word3* | | *K-translation3* |
| *word4* | *S-translation4* | |
| *word5* | | |

Table 4: English-Sorani and English-Kurmanji dictionaries structure

  The results would be structured as shown in Table 4.

- development of a transliteration/translation engine between Sorani and Kurmanji, that is capable of matching closely-related words and synstes.

- For the cases in which, more than one or no mapping has been found, manual filtering or insertion will be used.

### 4.4 Timing and Logistics

Based on our estimates, we plan to carry out the research highlighted in this paper in the course of one-and-an-half to two years. To this end, a timeline has been prepared (see Figure 5). We believe that since the preliminary work on KurdNet (e.g., literature review, development of the first prototype) has already been completed, most of our resources will be dedicated to designing new algorithms and system building.

Moreover, in terms of technical logistics, we are hopeful to receive full IT and library systems support from the Science and Research Branch Islamic Azad University(SRBIAU[12]) and University of Kurdistan(UoK[13]).

## 5 Summary

In this paper, we underlined the major shortcomings in the current KurdNet prototype and proposed a concrete plan to enrich the current prototype, so that it can be used in development of Kurdish language processing systems.

## Acknowledgment

The authors would like to express their gratitude to Yahoo! and Baidu for their generous travel and conference support for this paper.

[9] http://ckb.wikipedia.org/
[10] http://ku.wikipedia.org/
[11] http://multiwordnet.fbk.eu/
[12] http://krd.srbiau.ac.ir/
[13] http://www.uok.ac.ir/

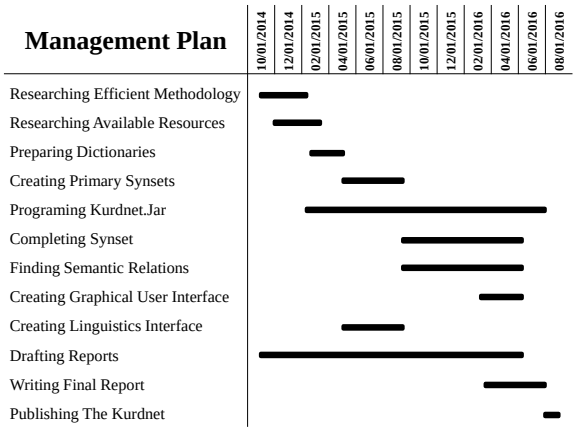


Figure 5: Management Plan

## References


Purya Aliabadi, Mohammad Sina Ahmadi, Shahin Salavati, and Kyumars Sheykh Esmaili. 2014. Towards Building KurdNet, the Kurdish WordNet. In *Proceedings of the 7th Global WordNet Conference (GWC'14)*, pages 1–6.

Richard Beckwith, George A. Miller, and Randee Tengi. 1993. Design and Implementation of the WordNet Lexical Database and Searching Software. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database*. MIT Press.

Pushpak Bhattacharyya. 2010. IndoWordNet. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'10)*.

Kyumars Sheykh Esmaili and Shahin Salavati. 2013. Sorani Kurdish versus Kurmanji Kurdish: An Empirical Comparison. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL'13)*, pages 300–305.

Kyumars Sheykh Esmaili, Shahin Salavati, and Anwitaman Datta. 2013a. Towards Kurdish Information Retrieval. *ACM Transactions on Asian Language Information Processing (TALIP)*, To Appear.

Kyumars Sheykh Esmaili, Shahin Salavati, Somayeh Yosefi, Donya Eliassi, Purya Aliabadi, Shownem Hakimi, and Asrin Mohammadi. 2013b. Building a Test Collection for Sorani Kurdish. In *Proceedings of the 10th IEEE/ACS International Conference on Computer Systems and Applications (AICCSA '13)*.

Christiane Fellbaum. 2010. *WordNet*. Springer.

Tiziano Flati and Roberto Navigli. 2012. The CQC Algorithm: Cycling in Graphs to Semantically Enrich and Enhance a Bilingual Dictionary. *Journal of Artificial Intelligence Research*, 43(1):135–171.

Gérard Gautier. 1998. Building a Kurdish Language Corpus: An Overview of the Technical Problems. In *Proceedings of ICEMCO*.

Goeffrey Haig and Yaron Matras. 2002. Kurdish Linguistics: A Brief Overview. *Language Typology and Universals*, 55(1).

Amir Hassanpour, Jaffer Sheyholislami, and Tove Skutnabb-Kangas. 2012. Introduction. Kurdish: Linguicide, Resistance and Hope. *International Journal of the Sociology of Language*, 217:1–8.

KLPP. 2013. KurdNet's Download Page. Available at: `https://github.com/klpp/kurdnet`.

David N. MacKenzie. 1961. *Kurdish Dialect Studies*. Oxford University Press.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. 1993. Introduction to WordNet: An On-line Lexical Database. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database*. MIT Press.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. Developing an Aligned Multilingual Database. In *Proceedings of the 1st Conference on Global WordNet (GWC'02)*.

Shahin Salavati, Kyumars Sheykh Esmaili, and Fardin Akhlaghian. 2013. Stemming for Kurdish Information Retrieval. In *The Proceeding (to appear) of the 9th Asian Information Retrieval Societies Conference (AIRS 2013)*.

Dan Tufis, Dan Cristea, and Sofia Stamou. 2004. BalkaNet: Aims, Methods, Results and Perspectives. A General Overview. *Romanian Journal of Information science and technology*, 7(1-2):9–43.

Piek Vossen, Laura Bloksma, Horacio Rodriguez, Salvador Climent, Nicoletta Calzolari, Adriana Roventini, Francesca Bertagna, Antonietta Alonge, and Wim Peters. 1998. The EuroWordNet Base Concepts and Top Ontology. *Deliverable D017 D*, 34:D036.

Piek Vossen. 1996. Right or Wrong: Combining Lexical Resources in the EuroWordNet Project. In *EURALEX*, volume 96, pages 715–728.

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

Géraldine Walther and Benoît Sagot. 2010. Developing a Large-scale Lexicon for a Less-Resourced Language. In *SaLTMiL's Workshop on Less-*