# Towards Building KurdNet, the Kurdish WordNet

**Purya Aliabadi**
SRBIAU
Sanandaj, Iran
purya.it@gmail.com

**Mohammad Sina Ahmadi**
University of Kurdistan
Sanandaj, Iran
reboir.ahmadi@gmail.com

**Shahin Salavati**
University of Kurdistan
Sanandaj, Iran
shahin.salavati@ieee.org

**Kyumars Sheykh Esmaili**
Nanyang Technological University
Singapore
kyumarss@ntu.edu.sg

## Abstract

In this paper we highlight the main challenges in building a lexical database for Kurdish, a resource-scarce and diverse language. We also report on our effort in building the first prototype of KurdNet – the Kurdish WordNet– along with a preliminary evaluation of its impact on Kurdish information retrieval.

## 1 Introduction

WordNet (Fellbaum, 1998) 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. Kurdish is a less-resourced language for which, among other resources, no wordnet has been built yet.

We have recently launched the Kurdish language processing project (KLPP[1]), aiming at providing basic tools and techniques for Kurdish text processing. This paper reports on KLPP's first outcomes on building KurdNet, the Kurdish WordNet.

At a high level, our approach is semi-automatic and centered around building a Kurdish alignment for Base Concepts (Vossen et al., 1998), which is a core subset of major meanings in WordNet. More specifically, we use a bilingual dictionary and simple set theory operations to translate and align synsets and use a corpus to extract usage examples. The effectiveness of our prototype database is evaluated via measuring its impact on a Kurdish information retrieval task. Throughout, we have made the following contributions:

1. highlight the main challenges in building a wordnet for the Kurdish language (Section 2),
2. identify a list of available resources that can facilitate the process of constructing such a lexical database for Kurdish (Section 3),
3. build the first prototype of KurdNet, the Kurdish WordNet (Section 4), and
4. conduct a preliminary set of experiments to evaluate the impact of KurdNet on Kurdish information retrieval (Section 5).

Moreover, a manual effort to translate the glosses and refine the automatically-generated outputs is currently underway.

The latest snapshot of KurdNet's prototype is freely accessible and can be obtained from (KLPP, 2013). We hope that making this database publicly available, will bolster research on Kurdish text processing in general, and on KurdNet in particular.

## 2 Challenges

In the following, we highlight the main challenges in Kurdish text processing, with a greater focus on

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

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Arabic-based | ا | ب | ج | چ | د | ێ | ف | گ | ژ | ک | ل | م | ن | ۆ | پ | ق | ر | س | ش | ت | وو | ڤ | خ | ز |
| Latin-based | A | B | C | Ç | D | Ê | F | G | J | K | L | M | N | O | P | Q | R | S | Ş | T | Û | V | X | Z |

(a) One-to-One Mappings

| | 25 | 26 | 27 | 28 |
|---|---|---|---|---|
| Arabic-based | ئ / ی | و | ی | ه |
| Latin-based | I | U / W | Y / Î | E / H |

(b) One-to-Two Mappings

| | 29 | 30 | 31 | 32 | 33 |
|---|---|---|---|---|---|
| Arabic-based | ڕ | ڵ | ع | غ | ح |
| Latin-based | (RR) | - | (E) | (X) | (H) |

(c) One-to-Zero Mappings

Figure 1: The Two Standard Kurdish Alphabets (Esmaili and Salavati, 2013)

the aspects that are relevant to building a Kurdish wordnet.

### 2.1 Diversity

Diversity –in both dialects and writing systems– is the primary challenge in Kurdish language processing (Gautier, 1998; Gautier, 1996; Esmaili, 2012). In fact, Kurdish is considered a *bi-standard*[2] 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. Figure 1 shows both of the standard Kurdish alphabets and the mappings between them.

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.

In summary, as the examples in (Gautier, 1998) show, the "same" word, when going from Sorani to Kurmanji, may at the same time go through several levels of change: writing systems, phonology, morphology, and sometimes semantics.

### 2.2 Complex Morphology

Kurdish has a complex morphology (Samvelian, 2007; Walther, 2011) and one of the main driving factors behind this complexity is the wide use of inflectional and derivational suffixes (Esmaili et al., 2013a). Moreover, as demonstrated by the example in Table 1, in the Sorani's writing system definiteness markers, possessive pronouns, enclitics, and many of the widely-used postpositions are used as suffixes (Salavati et al., 2013).

One important implication of this morphological complexity is that any corpus-based assistance or analysis (e.g., frequencies, co-occurrences, sample passages) would require a lemmatizer/morphological analyzer.

### 2.3 Resource-Scarceness

Although there exist a few resources which can be leveraged in building a wordnet for Kurdish – these are listed in Section 3– but some of the most crucial resources are yet to be built for this language. One of such resources is a collection of comprehensive monolingual and bilingual dictionaries. The main problem with the existing electronic dictionaries is that they are relatively small and have no notion of *sense*, *gender*, or *part-of-speech* labels.

Another necessary resource that is yet to be built, is a mapping system (i.e., a transliteration/translation engine) between the Sorani and Kurmanji dialects.

## 3 Available Resources

In this section we give a brief description of the linguistics resources that our team has built as well as other useful resources that are available on the Web.

### 3.1 KLPP Resources

The main Kurdish text processing resources that we have previously built are as follows:

– *the Pewan corpus* (Esmaili and Salavati, 2013): for both Sorani and Kurmanji dialects. Its basic statistics are shown in Table 2.

[2]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).

| دا | + | تان | + | یش | + | مکان | + | کتیو | = | کتیومکانیشتاندا |
|---|---|---|---|---|---|---|---|---|---|---|
| *daa* | + | *taan* | + | *ish* | + | *akaan* | + | *ktew* | = | *ktewakaanishtaandaa* |
| postpos. | + | poss. pron. | + | conj. | + | pl. def. mark. | + | lemma | = | word |

Table 1: An Exemplary Demonstration of Kurdish's Morphological Complexity (Salavati et al., 2013)

| | 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 2: The Pewan Corpus' Basic Statistics (Esmaili and Salavati, 2013)

– *the Pewan test collection* (Esmaili et al., 2013a; Esmaili et al., 2013b): built upon the Pewan corpus, this collection has a set of 22 queries (in Sorani and Kurmanji) and their corresponding relevance judgments.
– *the Payv lemmatizer*: it is the result of a major revision of Jedar (Salavati et al., 2013), our Kurdish *stemmer* whose outputs are stems and not lemmas. In order to return lemmas, Payv not only maintains a list of exceptions (e.g., named entities), but also takes into consideration Kurdish's inflectional rules.

### 3.2 Web Resources

To the best of our knowledge, here are the other existing readily-usable resources that can be obtain from the Web:
– *Dictio*[3]: an English-to-Sorani dictionary with more than 13,000 headwords. It employs a collaborative mechanism for enrichment.
– *Ferheng*[4]: 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).
– *Wikipedia*: it currently has more than 12,000 Sorani[5] and 20,000 Kurmanji[6] articles. One useful application of these entries is to build a parallel collection of named entities across both dialects.

## 4 KurdNet's First Prototype

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.

[3] http://dictio.kurditgroup.org/
[4] http://ferheng.org/?Daxistin
[5] http://ckb.wikipedia.org/
[6] http://ku.wikipedia.org/

### 4.1 Scope

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). The Kurmanji version will be built later and will be closely aligned with its Sorani counterpart. To that end, we have already started building a high-quality transliterator/translator engine between the two dialects.

### 4.2 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[7]. The Entity-Relationship (ER) model for the data represented in Base Concept is shown in Figure 2.

[7] http://globalwordnet.org/

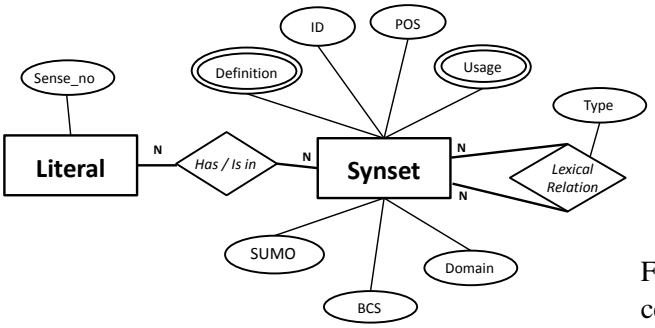



Figure 2: Base Concepts' ER Model

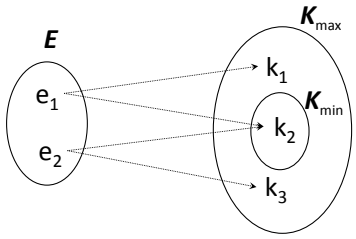



Figure 3: An Illustration of a Synset in Base Concepts and its Maximal and Minimal Alignment Variants in KurdNet

| | **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 3: The Main Statistical Properties of Base Concepts and its Alignment in KurdNet

### 4.3 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 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 3 shows an illustration of these two combination variants. In future, we plan to apply 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 Pewan's sentences (820,203), (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 must be aligned manually. The manual enrichment and assessment process is currently underway. We have built a graphical user interface to facilitate the lexicographers' task.

Table 3 shows a summary of KurdNet's statistical properties along with those of Base Concepts.

## 5 Preliminary Experiments

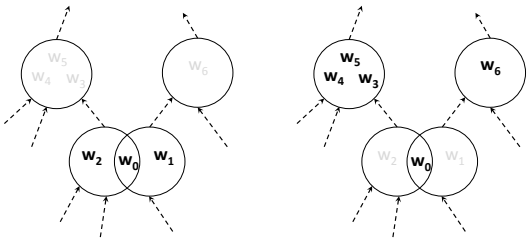
The most reliable way to evaluate the quality of a wordnet is to manually examine its content and structure. This is clearly very costly. In this paper we have adopted an indirect evaluation alternative in which we look at the effectiveness of using KurdNet for rewriting IR queries (i.e. query expansion).

We measure the impact of query expansion using two separate configurations: (i) **Terms**, which uses the raw version of the evaluation components (queries, corpus, and KurdNet), and (ii) **Lemmas**, which uses the lemmatized version of them. Furthermore, as depicted in Figure 4, we have considered two alternatives for expanding each query term: (i) add all of its **Synonyms**, and (ii) add all of the synonyms of its direct **Hypernym(s)**. Hence –given the *min* and *max* variants of KurdNet's synsets– there can be at least 10 different experimental scenarios.

In our experiments we have used the Pewan test collection (see Section 3.1), the **MG4J** IR engine (MG4J, 2013), and the Mean Average Precision (MAP) evaluation metric.

The results are summarized in Table 4. The notable patterns are as follows:

- since lemmatization yields additional




(a) By its Synonyms (b) By its Hypernyms

Figure 4: Expansion Alternatives for the Term $\mathbf{W}_0$

matches between query terms and their inflectional variants in the documents, it improves the performance (row 2 v. row 3). Expansion of the same lemmatized queries, however, degrades the performance (7-10 v. 1,4-6). This degradation can be attributed to the fact that the projection of KurdNet from terms to lemmas introduces imprecise entry merges.

- the *min* approach to align synsets outperforms its *max* counterpart overwhelmingly (1,4,7,8 v. 5,6,9,10), confirming the intuition that the *max* approach entails high-ambiguity,
- expanding query terms by their own synonyms is less effective than by their hypernyms' synonyms. This phenomena might be explained by the fact that currently for each query term, we use all of its synonyms and no sense disambiguation is applied.

Needless to say, a more detailed analysis of the outputs can provide further insights about the above results and claims.

## 6 Conclusions and Future Work

In this paper we briefly highlighted the main challenges in building a lexical database for the Kurdish language and presented the first prototype of KurdNet –the Kurdish WordNet– along with a preliminary evaluation of its impact on Kurdish IR.

We would like to note once more that the KurdNet project is a work in progress. Apart from the manual enrichment and assessment of the described prototype which is currently underway, there are many avenues to continue this work. First, we would like to extend our prototype to include the Kurmanji dialect. 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.

Another direction for future work is to prune the current structure i.e. handling the lexical idiosyncrasies between Kurdish and English.

| # | Scenario | MAP |
|---|---|---|
| 1 | `Terms & Hypernyms` (min) | *0.4265* |
| 2 | `Lemmas` | *0.4263* |
| 3 | `Terms` | *0.4075* |
| 4 | `Terms & Synonyms` (min) | *0.3978* |
| 5 | `Terms & Hypernyms` (max) | *0.3960* |
| 6 | `Terms & Synonyms` (max) | *0.3841* |
| 7 | `Lemmas & Hypernyms` (min) | *0.3840* |
| 8 | `Lemmas & Synonyms` (min) | *0.3587* |
| 9 | `Lemmas & Hypernyms` (max) | *0.2530* |
| 10 | `Lemmas & Synonyms` (max) | *0.2215* |

Table 4: Different KurdNet-based Query Expansion Scenarios and Their Impact on Kurdish IR

## References

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)*.

Kyumars Sheykh Esmaili. 2012. Challenges in Kurdish Text Processing. *CoRR*, abs/1212.0074.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

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. 1996. A Lexicographic Environment for Kurdish Language using 4th Dimension. In *Proceedings of ICEMCO*.

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.

MG4J. 2013. Managing Gigabytes for Java. Available at: `http://mg4j.dsi.unimi.it/`.

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)*.

Pollet Samvelian. 2007. A Lexical Account of Sorani Kurdish Prepositions. In *Proceedings of International Conference on Head-Driven Phrase Structure Grammar*, pages 235–249.

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-resourced Languages (LREC)*.

Géraldine Walther. 2011. Fitting into Morphological Structure: Accounting for Sorani Kurdish Endoclitics. In *The Proceedings of the Eighth Mediterranean Morphology Meeting*.