Developing a Large-Scale Lexicon for a Less-Resourced Language: General Methodology and Preliminary Experiments on Sorani Kurdish

Géraldine Walther[1] **& Benoît Sagot**[2]

1. LLF, Université Paris 7, 30 rue du Château des Rentiers, 75013 Paris, France
2. Alpage, INRIA Paris–Rocquencourt & Université Paris 7, Rocquencourt, BP 105, 78153 Le Chesnay Cedex, France
geraldine.walther@linguist.jussieu.fr, benoit.sagot@inria.fr

Abstract

In this paper, we describe a general methodology for developing a large-scale lexicon for a less-resourced language, i.e., a language for which raw internet-based corpora and general-purpose grammars are virtually the only existing resources. We apply this methodology to the development of a morphological lexicon for Sorani Kurdish, an Iranian language mostly spoken in northern Iraq and north-western Iran. Although preliminary, our results demonstrate the relevance of this methodology.

1. Introduction

Building large scale language resources for languages where there are only few linguistic resources and even less, if any, NLP resources available constitutes a challenge for NLP resource development. In this work, we aim at building a methodology which will allow us to develop new language resources for less-resourced languages from scratch. We will especially concentrate on the development of lexical resources, for the benefit they offer as such and as a starting point in the development other NLP resources and tools.

We first describe our methodology for building new language resources for resource-scarce languages (Section 3.). It uses solely raw on-line corpora and a few (basic) linguistic sources, such as simple reference grammars. In section 4., we illustrate this methodology with the description of SoraLex, a new, if preliminary, morphological lexicon for Sorani Kurdish destined to be enriched and completed with further syntactic information.[1]

2. Related work

In the past years, a large variety of approaches have been described aiming at developing morphological linguistic resources, in particular for less-resources languages. All of them try to benefit as much as possible from the limited amount of available data and information. Some approaches do not even rely on any prior linguistic knowledge, and fall in the paradigm of the unsupervised learning of a language's morphology (Goldsmith, 2001; Baroni, 2003; Creutz and Lagus, 2005). In such approaches, a raw corpus (usually in the form of a list of words) serves as the only input of the system, which automatically produces either a segmentation for each word into its morphemes, or even a full set of inflectional paradigms, associated with a set of lemmas (Snover and Brent, 2001). These techniques are useful for various purposes, including providing linguistic insights which are independent from the grammatical tradition of the considered language, if any. However, given the complexity and richness of morphological studies accessible for a very large range of languages, we agree with Forsberg et al. (2006) that it is time- and precision-wise counter-productive to try and automatically reproduce all this complexity instead of formalising morphological analyses available through linguistic literature.

In that regard, our approach is closer to most large-scale morphological resource development efforts (Ide and Véronis, 1994; Zanchetta and Baroni, 2005; Sagot, 2010), that also rely on explicit or implicit formalised morphological descriptions embedded in or compiled into part-of-speech (POS) taggers, lemmatisers and/or morphological analysers. However, we do not want to mandatorily rely on a lemmatiser or even on a POS tagger, as we aim at dealing with languages for which such tools do not yet exist. In further stages of the lexicon development, it shall of course become possible to POS-annotate a corpus of increasing size and therefore to train a POS-tagger, that shall give us access to acquisition techniques such as described by Molinero et al. (2009). However, we first need techniques that are able to automatically extract new lexical entries (i.e., lemmas and their associated inflection class), starting from a raw corpus and a formalised morphological description.

Several algorithms have been designed to extract new lemmas from such a limited amount of information. They have been applied to several languages such as Russian (Oliver et al., 2003), French verbs (Clément et al., 2004), German nouns (Perera and Witte, 2005), Slovak (Sagot, 2005), French verbs, nouns and adjectives (Forsberg et al., 2006) and Polish (Sagot, 2007). These techniques differ from one another in various aspects, such as the soundness of the underlying probabilistic model and/or heuristics, the richness of the manually described linguistic clues that are exploited (constraints on possible stems for each inflectional class, derivation patterns…), the use of Google for checking the "existence" of a form, or the possibility to benefit from (probabilised since uncertain) part-of-speech information when it becomes available.

In this work, we try to combine some of these ideas and techniques so as to define an efficient methodology for

[1] As we shall explain below, we call a *morphological lexicon* a set of entries of the form *(lemma,inflection class)* and the associated formalised description of the inflection classes. This allows for building, e.g., inflection and lemmatisation tools and a full-form lexicon (see below).

Kepa Sarasola, Francis M. Tyers, Mikel L. Forcada (eds.)
7th SaLTMiL Workshop on Creation and Use of Basic Lexical Resources for Less-Resourced Languages, LREC 2010, Valetta, Malta, 23 May 2010, p. 45–52

the development of a morphological lexicon for resource-scarce languages, and apply it to Sorani Kurdish.

3. A methodology for developing basic language resources from scratch

3.1. Constructing the morphological architecture

The most basic and yet most needed step in our language resource development is the construction of a morphological lexicon. A morphological lexicon associates a lemma and a morphosyntactic tag with each known wordform (form, in short).[2] However, building a morphological lexicon of a given language cannot be efficiently done without sufficient insight into this language's morphology. One needs to have at least access to a basic set of lexical entries and their morphosyntactic features in order to define the lemmas and the possible morphosyntactic tags of a given form. Our methodology therefore requests a preliminary study of the language's morphological specificities. These can however be extracted quite easily from simple linguistic descriptions of the language.

A summary linguistic study of the language's morphology should allow for the definition of a list of parts-of-speech together with their inflectional properties. From there, the linguistic descriptive features can be converted into an NLP tool-accessible language.

We chose to use the *Alexina* framework (Sagot, 2010) as a baseline for our lexical resource development. One asset of this framework lies in covering both the morphological and the syntactic level (e.g., valency) of a given lexicon — which shall be useful in further stages of the lexical resource development. *Alexina* offers an opportunity for representing lexical information in a complete, efficient and readable way (Sagot, 2005; Sagot, 2007; Sagot, 2010). Moreover it is compatible with the LMF standard[3] (Francopoulo et al., 2006).[4]

The *Alexina* model is based on a representation that separates the description of a lexicon from its use:

- The intensional lexicon factorises the lexical information by associating each lemma with a morphological class (previously defined in a formalised morphological description) and deep syntactic information; it is used for lexical resource development;

- The extensional lexicon, which is generated automatically by *compiling* the intensional lexicon, associates each inflected form with a detailed structure that represents all its morphological and syntactic information; it is directly used by NLP tools such as parsers.

Within this model, the necessary tasks for developing an intended new resource therefore consist in elaborating a formalised description of the targeted language's morphology, converting this description into the *Alexina* morphological language (Sagot, 2007) and finding possible lexical entries that can be associated with the inflection tables defined within the chosen *Alexina* model.

In the *Alexina* formalism, inflection is modelled as the affixation of a prefix and a suffix around a stem, while *sandhi* phenomena may occur at morpheme boundaries, sometimes conditioned by stem properties.[5] The formalism, which shares some widespread ideas with the DATR formalism (Evans and Gazdar, 1990) , relies on the following scheme:

- The core of a morphological description is a set of inflection classes which can (partly or completely) inherit from one another,

- Each inflection class defines a set of forms, each one of them being defined by a morphological tag and by a prefix and a suffix that, together with the stem, constitute the morpheme-like sequence *prefix_stem_suffix*;

- *Sandhi* phenomena allow to link the surface form to the underlying *prefix_stem* and *stem_suffix* sequences by applying regular transformations;

- Forms can be controlled by tests over the stem (e.g., a given rule can apply only if a given regular expression matches the stem and/or if another one does not match the stem);

- Forms can be controlled by "variants" of the inflection classes (e.g., forms can be selected by one or more flags which complement the name of the class).

Tables 2 and 1 in Section 4.3. illustrate this model by showing respectively a few *sandhi* rules and an excerpt of a verbal inflection class.

Translating a morphological description into the *Alexina* morphological language requires making choices about what will have to be treated as a dependent affix (prefixed or suffixed to the to-be-determined stems), an independent though typographically joined form or a typographically autonomous form. For that reason, the first descriptive task of the resource development consists in identifying the different affixes that can combine with possible stems. These identified affixes are used for constructing the inflectional tables associated with each of the previously defined inflectable parts-of-speech.

The second task consists in *somehow* gathering possible lexical entries for each part-of-speech (see Section 3.2.).

[2] Of course, a same form may receive more than one *(lemma,morphosyntactic tag)* pair.

[3] The Lexical Markup Framework ISO/TC37 standard.

[4] A fair number of lexical resources are already being developed within the *Alexina* framework, such as the Le*fff* for French (Sagot, 2010), the Le*ff*e for Spanish and other resources for Galician, Polish, Slovak, Persian and English. This fact should ensure the workability of new *Alexina* lexicons. It may also pave the way for future cross-language NLP applications.

[5] A *sandhi* — the term comes from traditional Sanskrit grammars — is a transformation of a given phonological/typographic sequence due to its encountering another specific sequence. The term *sandhi* is however nowadays used mainly although not always in order to refer to transformations occurring at morpheme boundaries. For example, in French, when the suffix *-ons* (1st person plural) is juxtaposed to the stem *mang-* (*to eat*), a *sandhi* phenomenon occurs that causes the insertion of a *e*, thus producing the form *mangeons* (*(we) eat*).

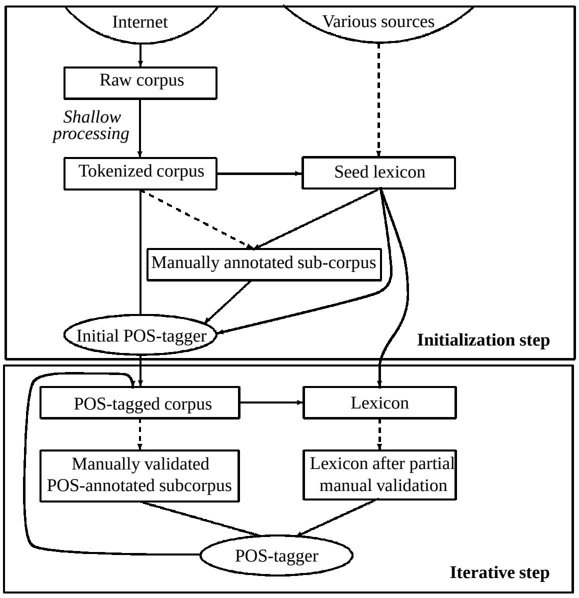

Figure 1: Overview of our lexical resource development methodology (dashed lines denote manual or semi-automatic steps, plain lines automatic steps)

3.2. Initialisation step: building a seed lexicon

Since an *Alexina* morphological lexicon consists of both a morphological description of a language and a set of lexical entries sorted according to their parts-of-speech, the next step in the development of the intended new lexical resource consists in finding possible candidates for the different word classes. This corresponds to the "initialisation step" in Figure 1.

To do so, one can either manually list a certain amount of lemmas associated with their inflectional class — that is, if there exists a resource listing such candidates — or, if such resources fail to be available, use the previously elaborated morphological description to infer the possible entries from obvious stems having combined with the identified affixes. All that is needed with this second method is a relatively large raw corpus. First we have to tokenise the raw corpus in order to extract a list of possible combinations of stems and affixes.[6] The types of possible combinations offer relatively accurate evidence for the classification of the inflectable words. This however only works with languages that display sufficiently rich inflectional classes and with those words which in fact combine with established affixes. For other cases, manual listing seems unavoidable.

After listing the lexical entries, we should be in possession of a small seed lexicon which will constitute the baseline for the development of all further large-scale resources.

3.3. Iterative step: enriching the lexicon and developing further resources

The "iterative step" of our methodology consists in using the newly built morphological seed lexicon to create other NLP tools which conversely allow to further develop the lexicon. The lexicon and these tools therefore benefit from each others' improvement.

Together with a limited-size manually POS-annotated corpus,[7] the information within the seed lexicon allows for building a specific lexicon-aware POS-tagger such as MElt (Denis and Sagot, 2009) for the newly-to-be-endued language. Once trained, MElt will be able to generate POS-tagged corpora for the targeted language, hence paving the way for the automatic extraction of candidate lexical entries thanks to simple techniques such as those described by Molinero et al. (2009). Of course, the newly suggested entries will require some (partial) manual validation, yet, since validating lexical entries is much less time-consuming than validating tagged corpora (Denis and Sagot, 2009) or, even more so, manually annotating raw corpora, this method does undoubtedly provide a much faster means for developing large-scale lexical resources from scratch.

A further expansion of the obtained resource would be the addition of the syntactic level of the lexicon, for which *Alexina* is already fully equipped. This step will require some more specific linguistic analysis and formalisation of the language's syntactic features, yet the necessary study of those features will conversely benefit from the existence of new POS-tagged corpora. Once the syntactic module of the new lexicon completed, it will also become possible to develop other NLP tools, such as parsers, for this language.

In brief, using the newly trained MElt-based POS-tagger will rapidly provide us with vast POS-tagged and morphologically annotated corpora, which will help improving the morphosyntactic lexicon, the underlying linguistic descriptions, and all other derived NLP tools. Thus, at any stage of our resource development, the interplay of the different modules enables an automatic iterative enrichment of each one of them.

4. A real case-study: SoraLex

We tested the above described methodology by building SoraLex, a morphological lexicon for Sorani Kurdish. For now, SoraLex only takes into account the morphological level of the intended lexicon, but it is destined to later be completed for syntactic information as well.

Sorani Kurdish is a resource-scarce language for which the only NLP resource available on the Internet seems to be raw text; as opposed to numerous other languages, there appear to be no usable on-line NLP tools accessible. We were therefore forced to build the whole development of our resource solely on some raw on-line corpora and the few existing linguistic descriptions of the language.[8] Since the first known description of the Kurdish language by Maurizio Garzoni (1787), only few other grammars have been published, none of them adopting a contemporarily formalised approach. For our description, we have been relying mainly on the descriptions of (McCarus, 1958; MacKenzie, 1961; Blau, 2000; Thackston, 2006).

Using solely those sources, we were yet able to build a preliminary version of the SoraLex morphological lexicon.

[6] We shall see in section 4. that the definition of the tokens sometimes requires a set of preliminary word vs. affix definitions.

[7] Say, a few hundreds of sentences.

[8] Although most of these descriptions are available on-line in the form of PDF documents, they can obviously not be considered as NLP resources and be used as such.

4.1. Sorani Kurdish language in brief

The Kurdish languages belong to the western branch of the Indo-Iranian languages. Kurdish speakers are mostly to be found in central and western Turkey, southern Iraq and Syria and western Iran. Yet a great number of Kurdish speakers also dwell in the neighbouring territories as well as spread all over the globe wherever the Kurdish diaspora has fancied to scatter them.[9]

Kurdish is composed of several dialects, of which the two major groups are the northern *Kurmanji*[10] written with extended Latin characters and the central/southern *Sorani*[11] written within a modified version of the Arabic script. Kurmanji and Sorani both possess a standardised form,[12] which, in the case of Sorani, has been largely shaped through the influence of the Kurdish Academy in Baghdad in the 1970s.

Standard Sorani Kurdish tends towards the Sulaymaniyah dialect of the north-eastern Iraqi province of As-Sulaymaniyah counting about one million Kurdish speakers.

Both in the remainder of this paper and in the SoraLex lexicon, we use an extension of the bijective transliteration employed for developing the PerLex lexicon for the Persian language (Sagot and Walther, 2010).[13]

4.2. The major morphological features

Sorani grammars (like those of the employed reference grammars) generally list the following parts-of-speech: nouns, verbs, pronouns and several *particles*.[14]

In our morphological description, we distinguish proper nouns, determiners, conjunctions, complementisers and prepositions in addition to the above mentioned classes. Though not yet explicitely linguistically motivated, our choices are preliminarily derived from usual classes within typological approaches. While most of these parts-of-speech correspond to their usual definition, the particle-class requires a closer look. Among the particles, we have counted the several pre- and postverbs (MacKenzie, 1961), adverbial suffixes and the second elements of Sorani circumpositions (*-ewe* and *-da*) (Thackston, 2006).

Concerning inflectional morphology, Sorani Kurdish, like most Indo-European languages, displays two major inflectional classes, the nominal class (including nouns, proper nouns, pronouns and adjectives) and the verbal class. In our approach towards the construction of a new lexical resource for Sorani Kurdish, those two classes have been endowed with a complete morphological description which has afterwards been adapted to the *Alexina* morphological language.

Concerning nominal inflection, the following elements have been included as affixal elements: the indefinite marker *-ěk*, the singular and plural definite marker *-eke*, and *-ekan*, the enclitic particle *-y* for marking modified nouns, called Ezafe, the enclitic pronominal person markers *-m*, *-t*, *-y*, *-man*, *-tan* and *-yan*, the demonstrative circumfixal demonstratives (Thackston, 2006) composed of the close *em-* and distant *ew-* respectively combined with the suffix *-e* and the focus particle *-yš*. As opposed to a certain amount of other Kurdish languages, Sorani Kurdish has lost any kind of case opposition between direct and oblique nominal forms, nor does it display any inflection for gender.

Other affixal elements linked to the nominal part of the Sorani Kurdish inflectional system are the comparative *-tar* and superlative *-taryn* attaching to adjectives only.

These different affixal markers can combine with each other, thereby creating rather complex morphological inflection pattern.

Still, future work shall aim at further defining the status of the Ezafe, the indefinite and definite markers and the enclitic personal suffixes, since their morphosyntactic properties clearly indicate a rather ambiguous status of these elements.

Concerning the verbal class, Sorani Kurdish resembles most Iranian languages in the fact that it possesses only a very limited amount of verbal lexemes (around 300). Most verbal meanings known from the more extensively described Indo-European languages are expressed through complex verbal predicates build from a light verbal head and a predicative element which can be either a noun or an adjective, or even an adposition or a pre- or postverb (MacKenzie, 1961; Blau, 2000).

The construction of Sorani verb forms obeys the following rules. Most descriptions concur in stating the existence of two distinct verbal stems, one (SI) for the present tense forms, one (SII) for the past tense forms.[15] For now, we also adopt this approach to Sorani verb morphology.

Sorani verb forms consist of the combination of a given stem with a set of pre- and suffixes, such as in the following representation:

Modal/Temporal Prefix(es) - Stem - Personal Suffix(es).

However, number of other elements may be inserted between the affixes and the stem. Enclitic pronominal person markers, for example, often appear between the modal prefix and the stem. Those specific difficulties yet have to be taken into account for SoraLex.

Sorani Kurdish also displays three sets of personal suffixes, the first being used with present verb forms derived from SI, the second with most past tense verb forms from SII

[9] In Europe, for example, a significant part of the important Turkish community is in fact of Kurdish origin.

[10] About 50% of the Kurdish speakers.

[11] About 25% of the Kurdish speakers.

[12] Established orthographic rules, standard uses, available normalised on-line corpora like newspapers and other websites.

[13] The use of a transliteration has at least two motivations. First, it allows for an easier development (e.g., text editors are not always very left-to-right-script-friendly, and lexicographers are not always familiar with the arabic script). Second, we use a latin-2 transliteration, which is compatible with NLP tools that require 8-bit encodings.

[14] Yet those lists appear to be incomplete and do not make the linguistic choices underlying the classification explicit. This part of Sorani linguistic description yet needs to be done.

[15] However this statement is not followed by (Bonami and Samvelian, 2008) who suggest the existence of three distinct stems, one for the present tenses, one for the past tenses and one for the passive forms. Our reading of the data contained within the reference grammars also gave us the impression that the question of the number of stems still needs to be solved. Depending on the question's outcome, the here presented morphological lexicon might yet expect some substantial changes.

and the third, being identical to the enclitic present forms of the verb *bwwn* ’to be’, with the remaining verb forms.

Yet the above-mentioned enclitic pronominal person markers may also function as agent-verb agreement markers for transitive verbs in the past tense. For those verbs, the normal personal suffixes function as patient markers. The interplay between the normal personal suffixes and the patient markers being a particularly complex phenomenon in Sorani Kurdish, we decided not to take into account the role of the pronominal person markers within the inflectional verb paradigm at this stage of our resource development and to wait for further linguistic insight. Concerning the pronominal person markers, linguistic motivation for treating them either within the morphological or the syntactic level of our *Alexina* lexicon might result in a substantial modifications in our morphological formalisation.

Having in mind the above sketched linguistically motivated morphological description, we have built a preliminary version of the morphological module of SoraLex.

4.3. An *Alexina* morphological description of Sorani Kurdish

As explained above, the first step of the development of the morphological module of SoraLex consisted in converting our morphological description gathered within the reference grammars of Sorani Kurdish so as to make them usable within the *Alexina* framework. Examples thereof are illustrated by the intransitive verb inflection class shown in Table 1 and in the noun inflection class shown in Table 5 together with a few sandhi rules in Table 2.[16]

```
<table name=”v1intrans” canonical_tag=”Inf”
                stems=”..*[aywdt]”>
    <form suffix=”n” tag=”Inf”/>
    <alt>
        <form suffix=”ww” tag=”PastPart” var=”c”/>
        <form suffix=”w” tag=”PastPart” var=”v”/>
    </alt>
. . .
    <form prefix=”de” suffix=”①m” tag=”1sgPreInd”/>
    <form prefix=”de” suffix=”①y” tag=”2sgPreInd”/>
    <form prefix=”de” suffix=”①ě” tag=”3sgPreInd”/>
    <form prefix=”de” suffix=”①yn” tag=”1plPreInd”/>
    <form prefix=”de” suffix=”①n” tag=”2plPreInd”/>
    <form prefix=”de” suffix=”①n” tag=”3plPreInd”/>
. . .
```

Table 1: Excerpts of the inflection class for Sorani Kurdish regular intransitive verbs in our *Alexina* morphological description.

Let us take the examples of the verbs *čwwn* ’to go’ and *parastn* ’to protect’. *Čwwn* belongs to the so called regular intransitive verbs shown in Table 1 which form their present stems by simply dropping their final vowel, whereas *parastn* counts as an irregular (transitive) verb, showing notably a case of vowel alternation between SI and SII. Their respective present and past stems are shown in Table 3.

```
<sandhi source=”ww_①” target=”_”/>
<sandhi source=”parast_①” target=”parěz_”/>
```

Table 2: A few *sandhi* rules from our *Alexina* description of Sorani Kurdish morphology, used to model the alternations between stems (the “_” models a morpheme boundary)

Canonical form	**Inflection class**	**SI**	**SII**
čwwn	**v1intrans**	*č-*	*čww-*
parastn	**v2trans**	*parěz-*	*parast-*

Table 3: Two verbal entries with their corresponding stems

Table 1 shows how the canonical form for intransitive verbs, the infinitive, is formed by adding the suffix *-n* (suffix=”**n**”) to the default stem SII. In fact, this also applies to transitive verbs. The past participle forms are similarly formed by adding either *-ww* or *-w*, depending on whether the stem ends respectively in a consonant (var=”**c**”) or a vowel (var=”**v**”), which is specified as a variant of the inflection class (“**v1intrans:v**” in the case of *čww-*, “**v2trans:c**” in the case of *parast-*). The present indicative forms make use of the sandhi phenomena shown in Table 2. Whenever the default stem (i.e., SII) encounters the symbol ① in an inflection table, the appropriate sandhi is triggered and the corresponding SI is generated. This results in the inflected forms shown in Table 4.

The case of nouns, illustrated in Table 6, is simpler, since nouns do not show stem alternations. Depending on the ending of their stems, they may only take certain forms of the following suffixes. As above, this constraint is modelled as inflection class variants (var=”**c**” for stems ending in consonants and var=”**v**” for stems ending in vowels).

Moreover, the “rads” and “except” constraints allow for further constraining the possible stems on which a suffix may attach: rads=”**.*[eěao]**” allows for the suffix to attach on any stem ending in *e*, *ě*, *a* or *o*, while except=”**.*[eěao]**” allows for the attaching of a given suffix to any stem except those ending in *e*, *ě*, *a* or *o*.

Table 6 shows an excerpt of the inflected forms for the nouns *dost* ’friend’ and *dě* ’village’, ending respectively in a consonant and in a vowel, as generated by the inflection class shown in Table 5.

4.4. Creation of a raw corpus and a seed lexicon

As mentioned above, the only other source of information we exploited is a raw corpus of Sorani Kurdish. We extracted such a corpus from the blog[17] of the programme *Ruwange* broadcasted by the Belgium-based Kurdish channel *Roj TV*. This blog allows for the automatic recursive retrieval of its pages, which we performed with the standard tool `wget`. We extracted all textual sections from the HTML files, removed all markup, filtered out lines that did not have the appearance of valid Sorani text (character set, spacing characteristics. . .) and segmented

[16] These tables are of course only excerpts of the full inflection tables contained within our *Alexina* description.

[17] `http://ruwange.blogspot.com/`

Inflected form	***čwwn***	***parastn***
Inf	*čww_n*	*parast_n*
PastPart	*čww_w*	*parast_ww*
1sgPreInd	*č_m*	*parěz_m*
2sgPreInd	*č_y*	*parěz_y*
3sgPreInd	*č_ě*	*parěz_ě*
1plPreInd	*č_yn*	*parěz_yn*
2plPreInd	*č_n*	*parěz_n*
3plPreInd	*č_n*	*parěz_n*

Table 4: Several inflected forms for the verbal entries in Table 3

it automatically into sentences based on final punctuation marks. Then we normalised[18] and transliterated all characters. We tokenised the corpus,[19] resulting in 590,568 token occurrences and 62,993 unique tokens. The most frequent tokens are the preposition *le*, the conjunction *w* and the preposition *be*.

With the help of this frequency list and the grammars listed above, we manually created a set of closed-class entries (29 conjunctions and complementisers, 22 punctuation marks, 10 determiners, 49 prepositions, 26 pronouns, 38 numerals, 10 particles). We also built a lexicon of 68 verb lemmas, which already covers almost 25% of the full set of Sorani Kurdish verbs.

In order to extract nouns, adjectives and adverbs from our corpus in a more systematic way, we decided to start with a simple technique, based on our knowledge of Sorani Kurdish morphology. We designed a regular expression[20] covering a large range of possible nominal and adjectival suffixes, such that the removal of these suffixes provides a nominal or adjectival candidate stem, i.e., in Sorani Kurdish, a lemma. In order to rank the obtained lemmas, we take advantage of the following hypotheses. First, the longer a suffix, the more probable it is correctly identified, and therefore its removal provides a valid nominal or adjectival lemma. Second, the more different suffixes have been identified on a given stem/lemma, the more confident we are in its correctness. Therefore, we assigned to each suffix a weight equal to its length, and weighted each candidate lemma by the sum of the weights of all (unique) suffixes it has been encountered with. This resulted in a list of 1,009 candidate lemmas with a weight of 10 or more, for which we performed a partial manual validation.

```
<table name=”N1” rads=”..*”>
    <form suffix=”” tag=”Abs”/>
    <alt>
        <form suffix=”ěk” tag=”SingIndef”
                rads=”.*” var=”c”/>
        <form suffix=”ě” tag=”SingIndefFam”
                rads=”.*” var=”c”/>
        <form suffix=”yěk” tag=”SingIndef”
                rads=”.*” var=”v”/>
        <form suffix=”yě” tag=”SingIndefFam”
                rads=”.*” var=”v”/>
        <form suffix=”yek” tag=”SingIndefFam”
                rads=”.*” var=”v”/>
        <form suffix=”ye” tag=”SingIndefFam”
                rads=”.*” var=”v”/>
    </alt>
...
    <alt>
        <form suffix=”an” tag=”PlIndef”
                rads=”.*” var=”c”/>
        <form suffix=”yan” tag=”PlIndef”
                rads=”.*” var=”v”/>
    </alt>
...
    </alt>
        <form suffix=”eke” tag=”SingDef”
                except=”.*[eěao]” var=”c”/>
        <form suffix=”eke” tag=”SingDef”
                except=”.*[eěao]” var=”v”/>
        <form suffix=”ke” tag=”SingDef”
                rads=”.*[eěao]” var=”v”/>
    </alt>
...
    <alt>
        <form suffix=”ekan” tag=”PlDef”
                except=”.*[eěao]” var=”c”/>
        <form suffix=”ekan” tag=”PlDef”
                except=”.*[eěao]” var=”v”/>
        <form suffix=”kan” tag=”PlDef”
                rads=”.*[eěao]” var=”v”/>
    </alt>
...
```

Table 5: Excerpts of the inflection class for Sorani Kurdish nouns in our *Alexina* morphological description

Inflected form	***dost***	***dě***
SingIndef	*dost_ěk*	*dě_yěk*
SingIndefFam	*dost_ě*	*dě_yě* *dě_yek* *dě_ye*
PlIndef	*dost_an*	*dě_yan*
SingDef	*dost_ke*	*dě_ke*
PlDef	*dost_ekan*	*dě_kan*

Table 6: Several inflected forms for the nouns *dě* ’village’ and *dost* ’friend’

In order to build additional open-class candidates, we also applied our implementation of the algorithm described in (Sagot, 2005). This algorithm is based on the list of unknown and open-class tokens associated with their frequencies. On our corpus, and taking into account

[18] Sorani Kurdish, as Urdu, has the following property. The isolated and final forms of the Arabic letter *hâ* constitute one letter (pronounced *e*), whereas the initial and medial forms of the same Arabic letter constitute *another* letter (pronounced *h*), for which a different Unicode encoding is available. In many electronic texts, such as the blog we used as a corpus, these letters are written using only the *hâ*, and differentiate both letters using the *zero-width non joiner* character that prevents a character from being joined to its follower. We had to normalise this in order to get two different Unicode encodings for these two different letters.

[19] For this task we used a simple tokeniser, that only recognises numbers, URLs, email addresses and a few other very surfacic phenomena. It then identifies all punctuation marks as individual tokens, as well as all remaining sequences of non-whitespace characters.

[20] `((y[eě]|ě)(k(an|e)?)?)?(y?š)?([mty](an)?)?y?$`

the already existing entries, we obtained 4,104 candidate lemmas, ordered according to a weight that takes into account both the likelihood of each lemma as computed by the algorithm and the number of occurrences of its inflected forms. We manually validated a limited amount of these candidates. A web-based interface already developed and used for other lexical development projects shall allow for an efficient large-scale manual validation of these candidate entries, and therefore improve the coverage of SoraLex in the near future.

Finally, we used the Sorani Kurdish Wikipedia[21] for collecting proper nouns. Those were found through the titles of Wikipedia articles indicating either a city, a country or a person *category*. We collected and normalised the titles of these articles as well as those of all the articles redirecting towards them. We were thereby able to build a lexicon for proper nouns consisting in person, country and city names. These tasks resulted in a set of (only) 131 proper noun lemmas. Person names have been assigned the class of invariable lemmas, whereas countries and cities received an inflectional noun class that doesn't allow for the formation of plural forms.

Using these manual and semi-automatic techniques, we obtained a seed lexicon for Sorani Kurdish. This lexicon contains 17,600 extensional (form-level) entries corresponding to 13,315 different forms from 468 intensional (lemma-level) entries. This lexicon covers 48.4% of all token occurrences in our raw corpus.

5. Conclusion

In this paper, we introduced a three-step methodology for developing morphological lexicons for resource-scarce languages, i.e., languages for which raw corpora and linguistic studies are basically the only available sources of information. First, we argued for the relevance of a careful linguistic study allowing for the manual development of a formalised description of the language's morphology. In a second step, the initialisation step, we suggest employing both existing and novel techniques that use such a description for constructing semi-automatically a *seed* lexicon from a raw corpus of the language. Coupled with a (small) manually annotated corpus, this seed lexicon helps training a preliminary version of a lexicon-aware part-of-speech tagger such as MElt (Denis and Sagot, 2009), which enables to generate a large POS-tagged corpus. Such a corpus is in turn useful for efficiently improving the coverage of the lexicon (Molinero et al., 2009), and therefore the quality of the tagger, thus defining a virtuous iterative process.

We illustrate this methodology by reporting the first steps towards the development of a large-scale morphological lexicon for Sorani Kurdish within the *Alexina* framework. We are currently about to move from the initialisation step to the iterative step. Apart from following our methodology, we aim at exploring other complementary approaches. In particular, we plan to develop techniques for extracting relevant information from existing lexical resources available for closely related languages. Ongoing work in this direction has given satisfying results for the Galician language starting from resources for Spanish, and we intend to benefit from the ongoing initiative around the PerLex lexicon for Persian (Sagot and Walther, 2010) so as to try and gather complementary information.[22]

On the longer term, we intend to develop a first set of NLP tools for Sorani Kurdish based on SoraLex and existing technologies already adapted to Persian language based on PerLex. This includes, among others, advanced tokenisation and segmentation modules, named entity recognisers and spelling correctors.

SoraLex, as all *Alexina* lexicons, is available under a free software license (LGPL-LR) on the web-page of the *Alexina* project.[23]

6. References

Marco Baroni. 2003. Distribution-driven morpheme discovery: A computational/experimental study. In Geert Booij and Jaap van Marle, editors, *Yearbook of Morphology 2003*, pages 213–248. Dordrecht: Springer.

Joyce Blau. 2000. *Méthode de kurde sorani*. L'Harmattan, Paris, France.

Olivier Bonami and Pollet Samvelian. 2008. Sorani kurdish person markers and the typology of agreement. In *13th International Morphology Meeting*, Vienna, Austria.

Lionel Clément, Benoît Sagot, and Bernard Lang. 2004. Morphology based automatic acquisition of large-coverage lexica. In *Proceedings of the 4th Language Resources and Evaluation Conference (LREC'04)*, pages 1841–1844, Lisbon, Portugal.

Mathias Creutz and Krista Lagus. 2005. Inducing the morphological lexicon of a natural language from unannotated text. In *Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR'05)*, pages 106–113, Espoo, Finland.

Pascal Denis and Benoît Sagot. 2009. Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art POS tagging with less human effort. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation (PACLIC 2009)*, Hong Kong.

Roger Evans and Gerald Gazdar. 1990. The DATR Papers: February 1990. Technical Report CSRP 139, University of Sussex, Brighton, UK.

Markus Forsberg, Harald Hammarström, and Aarne Ranta. 2006. Morphological lexicon extraction from raw text data. In *Proceedings of FinTAL 2006, LNAI 4139*, pages 488–499, Turku, Finland. Springer-Verlag.

Gil Francopoulo, Monte George, Nicoletta Calzolari, Monica Monachini, Nuria Bel, Mandy Pet, and Claudia Soria. 2006. Lexical Markup Framework (LMF).

[21] Available at the following address: `http://ckb.wikipedia.org`. We used the dump of March 26th, 2010.

[22] This idea could also be used in order to develop a lexicon for Kurmanji Kurdish from SoraLex, and/or to benefit from existing limited-size lexical resources for this language.

[23] `http://alexina.gforge.inria.fr/`

In *Proceedings of the 5th Language Resources and Evaluation Conference (LREC'06)*, Genoa, Italy.

Maurizio Garzoni. 1787. *Grammatica e Vocabulario della Lingua Kurda*. Sacra Congregazione di Propaganda Fide, Rome, Italy.

John A. Goldsmith. 2001. Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, 27(2):153–198.

Nancy Ide and Jean Véronis. 1994. MULTEXT: Multilingual text tools and corpora. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING'94)*, Kyoto, Japan.

David N. MacKenzie. 1961. *Kurdish dialect studies*, volume 1 of *London Oriental Series*. Oxford University Press, London, UK.

Ernest N. McCarus. 1958. *A Kurdish Grammar: descriptive analysis of the Kurdish of Sulaimaniya, Iraq*. Ph.D. thesis, American Council of Learned Societies, New-York, USA.

Miguel Ángel Molinero, Benoît Sagot, and Lionel Nicolas. 2009. A morphological and syntactic wide-coverage lexicon for Spanish: The Le*ffe*. In *Proceedings of the 7th conference on Recent Advances in Natural Language Processing (RANLP 2009)*, Borovets, Bulgaria.

Antoni Oliver, Irene Castellón, and Lluís Màrquez. 2003. Use of Internet for augmenting coverage in a lexical acquisition system from raw corpora: application to Russian. In *Proceedings of the RANLP'03 International Workshop on Information Extraction for Slavonic and Other Central and Eastern European Languages (IESL'03)*, Borovets, Bulgaria.

Praharshana Perera and René Witte. 2005. A self-learning context-aware lemmatizer for German. In *Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, pages 636–643, Vancouver, Canada.

Benoît Sagot and Géraldine Walther. 2010. A morphological lexicon for the Persian language. In *Proceedings of the 7th Language Resources and Evaluation Conference (LREC'10)*, Valetta, Malta. To appear.

Benoît Sagot. 2005. Automatic acquisition of a Slovak lexicon from a raw corpus. In *Lecture Notes in Artificial Intelligence 3658 ((c) Springer-Verlag), Proceedings of TSD'05*, pages 156–163, Karlovy Vary, Czech Republic.

Benoît Sagot. 2007. Building a morphosyntactic lexicon and a pre-syntactic processing chain for Polish. In *Proceedings of the 3rd Language & Technology Conference (LTC'05)*, pages 423–427, Poznań, Poland.

Benoît Sagot. 2010. The Le*fff*, a freely available, accurate and large-coverage lexicon for French. In *Proceedings of the 7th Language Resources and Evaluation Conference (LREC'10)*, Valetta, Malta. To appear.

Matthew G. Snover and Michael R. Brent. 2001. A bayesian model for morpheme and paradigm identification. In *Proceedings of the 39th annual meeting of the ACL*, pages 490–498, Toulouse, France.

Wheeler M. Thackston. 2006. Sorani kurdish: A reference grammar with selected readings. www.fas.harvard.edu/˜iranian/Sorani/sorani_1_grammar.pdf.

Eros Zanchetta and Marco Baroni. 2005. Morph-it! a free corpus-based morphological resource for the Italian language. In *Proceedings of Corpus Linguistics 2005*, Birmingham, UK. University of Birmingham.