Fast Development of Basic NLP Tools: Towards a Lexicon and a POS Tagger for Kurmanji Kurdish

Géraldine Walther, Benoît Sagot, Karen Fort

Fast Development of Basic NLP Tools: Towards a Lexicon and a POS Tagger for Kurmanji Kurdish

Géraldine Walther[1], Benoît Sagot[2], Karën Fort[3]
1. Laboratoire de Linguistique Formelle, Université Paris 7
2. Alpage, INRIA Paris–Rocquencourt & Université Paris 7
3. Institut de l'information scientifique et technique (INIST), CNRS
geraldine.walther@linguist.jussieu.fr, benoit.sagot@inria.fr, karen.fort@inist.fr

Abstract. The development of basic NLP resources for minority languages is still a challenge to both formal and computational linguists. In this paper, we show how we were able to develop a medium-scale morphological lexicon for Kurmanji Kurdish in a few days time using only freely accessible resources. We also developed a preliminary POS tagger that shall be used as a pre-annotation tool for developing a POS-annotated corpus, based solely on raw text and on our morphological lexicon.

1 Introduction

Among the world's languages, even those vastly spoken, there exists a great proportion for which no usable NLP tools are yet available. In particular, no lexical resources are freely available, though such resources, together with basic pre-processing tools such as POS taggers, constitute the basis for the development of more complex NLP applications, such as parsing, machine translation, data-mining or information extraction. Moreover they also directly benefit linguists who wish to study these languages in more detail and therefore need vast reliably tagged corpora.

In this paper, we show that it is possible to build, in quite a short time, a range of basic NLP resources, namely a lexicon and a POS tagger, for a formerly unequipped language.

As an example we built a morphological lexicon and a fairly accurate POS tagger for Kurmanji Kurdish[1] using solely the following data which was easily accessible since freely available on the Internet: lexical information from various sources, a non-formalised descriptive reference grammar (Thackston, 2006) and a raw corpus of several thousand words.

Our work consisted in the following five steps:

1. We built an inventory of Kurmanji categories from information contained in the reference grammar. In parallel, we wrote a formalised description of Kurmanji morphology in the lexical formalism Alexina.
2. We then extracted lexical entries from various lexical information sources and inflected them thanks to the formalised morphological description mentionend in 1, thus obtaining KurLex, a morphological lexicon for Kurmanji Kurdish.
3. We designed a tagset containing 36 tags designed for POS tagging by refining the list of categories (we therefore use *POS* as a synonym for *tag* rather than *category*, for consistency with the term *POS tagger*).
4. We automatically generated a POS annotated corpus using only that lexical information.
5. We finally trained the MElt tagger (Denis & Sagot, 2009) on this corpus and on KurLex. Its 85.7% accuracy, obtained without a manually annotated corpus, is not yet suitable for actual tagging tasks. However, it has been proven sufficient as a pre-annotator for helping the manual development of a POS-annotated corpus for English (Fort & Sagot, 2010), with improvements over fully manual annotation in both time and quality that are comparable to the best results obtained by pre-annotating with a state-of-the-art POS tagger.

[1] Kurdish is a Western Iranian language mainly spoken in Eastern Turkey, but also in Northern Syria and Iraq, Western Iran and the Kurdish diaspora. Kurmanji is the most widely used variant of Kurdish (half of the 20 million Kurdish speakers, most of themcoming from the Turkish regions).

2 Creating a Formal Description of Kurmanji Kurdish Morphology

2.1 Building a Morphological Description for Kurmanji Kurdish

Our first task consisted in establishing a sound set of categories. Our Kurmanji reference grammar (Thackston, 2006) lists the following categories: nouns, verbs, pronouns, numerals, adjectives, pre-, post- and circumpositions, complementisers and several *particles*. In our morphological description, we do not take circumpositions into account since they always consist of a combination of a preposition and one of the three postpositional elements *de, re* and *ve*. Our choices are preliminarily derived from usual classes within typological approaches and partly linguistically motivated in the sense that they try to combine common basic morphological features and distributional constraints for each category.[2] In addition to the above mentioned classes we distinguish the following categories: proper nouns, determiners and conjunctions. Kurmanji Kurdish seems to have a very limited amount of words that might be called adverbs, most of them being a combination of an adjective preceded by the derivative particle *bi*. We therefore decided to not consider adverbs as an independent category. Some of the elements which could be considered as adverbs are listed as particles, others are considered as a combination of an adjective preceded by the derivative particle *bi*.

The full set of categories, used in our morphological description and in the lexicon, is described in table 1.

Our second task was to create a formalised description of Kurmanji Kurdish from information contained in (Thackston, 2006). Like most Indo-European languages, Kurmanji Kurdish displays two major inflectional classes, the nominal class (including nouns, proper nouns, pronouns and adjectives) and the verbal class. Those two classes have been endowed with a complete morphological description.

Kurmanji nouns and pronouns show different forms for case, number and gender. The singular and plural definite case endings of the masculine and feminine genders *-ê, -î, -ên, -an* and *-a* for the construct (*Ezafe*), oblique and demonstrative cases have been treated as affixal elements. The indefinite marker *-(y)ĕk*, which may be inflected for number, gender and case has been treated as an affixal element as well. Other affixal elements within the Kurmanji Kurdish nominal inflectional system are the comparative *-tir* and superlative *-tirîn* attaching to adjectives (and participles). Apart from that, adjectives do not show gender or number inflection. Concerning verbs, Kurdish, like most Iranian languages, possesses only a limited amount of verbal lexemes (about 300).[3] The construction of Kurmanji verb forms obeys the following rules. Most descriptions, including (Thackston, 2006) state the existence of two distinct verbal stems, one (SI) for the present tense forms and one (SII) for the past tense forms. The imperative forms are considered SI forms. Yet some verbs show a third imperative stem (SIII) for which we have made space in our fomalised morphological description. We therefore chose a three-stem description for representing Kurmanji verbs. All verb forms combine a given stem with a set of pre- and suffixes, such as in the following representation:

(Negation Prefix) — Temporal/Aspectual/Modal Prefix(es) – Stem – Personal Suffix(es).[4]

Kurmanji Kurdish displays a typical form of split ergativity (Creissels, 2006): in the present tenses the verb shows A-agreement with S-A alignment (Dixon, 1994), as do intransitive verbs in the past tenses, where transitive verbs show P-agreement with S-P-alignment.

---

[2] We are however aware that the description of the category underlying our NLP-tool development would by no means suffice in a theoretical analysis of Kurmanji Kurdish. It still requires explicit theoretical clarification to be linguistically satisfying.

[3] Most verbal meanings known from the more extensively described Indo-European languages are expressed through complex verbal predicates built from a light verbal head and a predicative element.

[4] There are two sets of personal suffixes, the first being used with present verb forms derived from SI (and SIII), the second with most past tense verb forms from SII.

CATEGORY	EXAMPLE	DESCRIPTION
PUNCT	, . ? ... ! ;	*punctuation signs*
N	HOZANTÎ — *FemDefSgNom "poetry"* EWRINAN — *MascIndefPlObl "cloud"*	*nouns* *inflecting in case, number, gender & defineteness*
PN	KURDISTAN — *"Kurdistan"*	*proper nouns (no number inflection)*
PRN	EZ — *PersP1nom "I"* MIN — *PersP1obl "I"* ÇI — *negative pronoun* XWE — *reflexive pronoun* HEV — *pronoun of reciprocity* KU/KO — *relative pronoun*	*pronouns* *inflecting in case, number, gender*
V	KIRIN — *InfAff "to do"* NEÇUYE — *P2|3sgNegIndPresPerf*	*verbforms* *inflecting for person, number, time, aspect, modality & negativity*
ADJ	DÛR — *Norm* "FAR" MEZINTIR — *Comp "bigger"* BEXTYARTIRÎN — *Super "luckiest"*	*adjectives* *inflecting for comparative and superlative*
NUM	ÇAR — *Card "4"* ÇARAN — *Ord "4th"*	*numerals* *inflecting for cardinality or ordinality*
PREP	BA *"to, towards"*	*prepositions (or first elements of traditional circumpositions)*
POSTP	VE, DE, RE | VA, DA, RA	*postpositions (or second element of traditional circumpositions)*
COMPL	GAVA (KU) — *"as long as"*	*complementisers*
PART	JÎ — *emphatic* ÇEND — *question "how much"* BI — *derivation* NE — *negation*	*particles*
DET	HER *"every"*	*determiners*
CONJ	Û — *"and"*	*coordinating conjonctions*

Table 1: Our inventory of Kurmanji Kurdish categories

2.2 Formalising Kurmanji Morphology within the Alexina Lexical Framework

We formalised these morphological features within the Alexina framework (Sagot, 2010), thus paving the way for an Alexina lexicon for Kurmanji Kurdish, named KurLex, alongside already existing lexicons such as the Le*fff* for French. One asset of this framework lies in covering both the morphological and the syntactic levels (e.g., valency) — which shall be useful in further stages of our work. Alexina offers an opportunity for acquiring and representing lexical information in a complete, efficient and readable way (Sagot, 2007; Sagot, 2010). Moreover it is compatible with the LMF (Lexical Markup Framework) ISO standard (Francopoulo *et al.*, 2006).

In the current 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 & Gazdar, 1990) and other similar work (Skoumalová, 1997), 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).

The KurLex Alexina description contains 8 inflection classes for verbs, 5 classes for nouns (among which one concerns proper nouns with singular forms only), 2 classes for each adjectives and determiners, 1 class for each numerals, pronouns and particles. All other lemmas are given one of the three invariable classes (*inv* for invariable lemmas and *pref* resp. *suff* for prefixes resp. suffixes).

3 Building the KurLex Lexicon

In order to produce a full-form lexicon (i.e., triples of the form *(form, lemma, morphological tag)*), we needed the largest lexicon of lemmas possible.

We first developed various scripts for converting the following lexical resources into the Alexina format, extracting as much information as possible:[6]

- 79 verbs with their three stems from a native speaker of Kurmanji (Öpengin, p.c.);
- the Kurmanji-English vocabulary from Thackston (2006) (almost 4,700 entries extracted);
- the glossary developed by the *Institut Kurde de Paris*[7] as a byproduct of the Kurmancî linguistic magazine, and made freely available (almost 6,700 entries extracted);[8]

[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*).

[6] Including the category and translations in English, when available. This availability can be indirect. For instance, an English translation of the title of a Wikipedia article can be retrieved thanks to inter-wiki links, and its category is `PN` (proper noun) if its Wikipedia categories indicate that they denote named entities such as cities, countries, persons etc.

[7] http://www.institutkurde.org/en/

[8] http://www.institutkurde.org/en/publications/kurmanci/downloads/

– the Open Office spell-checker lexicon (over 4,700 entries extracted);[9]
– the Kurdish Wiktionary (over 31,000 entries extracted);[10]
– the Kurdish Wikipedia (360 extracted article titles from specific categories — cities, countries, etc.);[11]

Next, we merged these lists of lemmas by computing an optimal (partial) mapping between entries from these various sources. For all pairs of entries with the same canonical form, we computed an equality likelihood, taking into account the category when available and comparable[12] as well as the relative overlap of the lists of English translations, when available. This resulted in a lexicon of 30,505 entries, out of which 25,228 with a known category.

We filtered this preliminary lexicon by removing entries that were obviously erroneous (incorrect language or script[13], very long entries that clearly correspond to idioms. . .).[14]

Finally, we generated the inflected version of the lexicon. We also compiled, with no additional work, a Kurmanji version of the SxPipe pre-processing chain. We applied it on the *selected readings* joint with Thackston's (2006) grammar (23,711 tokens) as a raw corpus, from which we excluded a small evaluation subcorpus (168 tokens, see below). This allowed us to identify and count unknown word occurrences. After manually adding a few frequent missing entries, our lexicon, named KurLex, contains 22,327 entries[15] that generate 412,320 inflected form entries for 235,280 unique inflected forms. Its coverage on our raw corpus is approximately 83%. KurLex is freely available as part of Alexina.[16]

4 Building the POS Tagger

4.1 Related work

A number of studies were conducted concerning the creation of a POS tagger with limited resources. Concerning the design of the tagset, one particular trend from machine translation involves the use of a "target" language information to semi-automatically generate the tagset (Cucerzan & Yarowsky, 2002; Sánchez-Martínez *et al.*, 2008). Other solutions include adding morphological information to improve tagset accuracy for a morphologically rich language (Dandapat *et al.*, 2007). In this work, we extended the set of category defined in section 2 with some morphological features extracted from KurLex (e.g., pronoun types, noun definiteness, etc.), resulting in a more detailed tagset. This tagset contains 36 POS (note that each POS, i.e., tag, corresponds to one of the categories listed in 1 or a refinement of one of them).[17] We generated the corresponding KurLex variant using these tags, to be used as a source of tagging information. The resulting *(form,tag)* lexicon is called $\text{KurLex}_{\text{tags}}$.

[9] See http://wiki.services.openoffice.org/wiki/Dictionaries

[10] Wîkîferheng, http://ku.wiktionary.org/

[11] Wîkîpediya, http://ku.wikipedia.org/

[12] E.g., some of our input resources distinguish between masculine, feminine and plural nouns, whereas others do not.

[13] This happened on Wiktionary data, as words in other languages and other varieties of Kurdish do have pages in the Kurdish Wiktionary.

[14] For now, we also filtered out regional or dialectal variants mentioned in Thackston's lexicon.

[15] These lemma-level entries include among others 16,953 common nouns, 3,727 adjectives, 1,091 proper nouns, 333 verbs, 41 numerals, 45 prepositions, 27 complementisers, 10 particles, 6 postpositions.

[16] https://gforge.inria.fr/projects/alexina/

[17] One punctuation tag, two noun tags (definite and indefinite), one proper noun tag, eight pronoun tags (four for personal pronouns respectively in the nominative, oblique, construct and demonstrative cases and one respectively for each negative, reflexive, reciprocal and relative pronoun), seven verb tags (for verb forms built from the past stem, the present stem, infinitive forms — in the nominative, the oblique or the construct case —, imperatives and participles) three adjective tags (basic, comparative and superlative forms), two numeral tags (cardinal and ordinal numbers), one preposition and one postposition tag, one complementiser tag, six particle tags (emphatic, question, secondary construct particle, derivative particle, negation and *others*), two derterminer tags (*her* "every" and demonstrative determiners) and one coordinating conjunction tag.

The POS tagger construction task proper is a variant of the unsupervised POS tagging problem. It differs from the purely unsupervised task in so far as a dictionary is available, that provides a list of allowable tags for each word it contains. This task has been studied by various authors, following the seminal work of Merialdo (see for example (Merialdo, 1994)). In most cases, machine learning techniques are used for inducing a probabilistic model of some kind (see however (Brill, 1995)). The most popular model for this task is HMMs (Merialdo, 1994; Goldwater & Griffiths, 2007; Goldberg *et al.*, 2008; Ravi & Knight, 2009), but other models have been succesfully proposed, such as discriminative models (Smith & Eisner, 2005). Various training methods have been used, starting with the standard Baum-Welsh expectation-maximisation algorithm, and ranging from a fully Bayesian approach (Goldwater & Griffiths, 2007) to the so-called contrastive-evaluation (Smith & Eisner, 2005) or sophisticated MDL-based techniques (Ravi & Knight, 2009). Linguistic knowledge is sometimes used, e.g., for initializing the parameters of the HMM model. Sometimes, ambiguity classes (the sets of part-of-speech tags a type can appear with) are explicitely modeled (Toutanova & Johnson, 2008). In this work, rather than improving the quality of a single model using sophisticated statistical techniques, we have tried a different approach. We developed several different models using various simple statistical approaches and heuristics, we created an automatically annotated corpus based on all these models using inter-model agreement measures, and we finally trained on this corpus a POS tagger known to achieve state-of-the art accuracy for supervised POS tagging.

Maximum-entropy taggers proved to be the best in several comparison experiments, for example on Amharic (Gambäck *et al.*, 2009), on Bengali, on small training data (Dandapat *et al.*, 2007), or on French (Denis & Sagot, 2009). Among these, MElt (Denis & Sagot, 2009), used in this work, has been designed with a specific effort on integrating lexical information, which is particulary relevant here.

4.2 Building a MElt tagger for Kurmanji Kurdish

In order to minimise the amount of manual (and expert) work, we tried to develop a training corpus for the MElt tagger based solely on the lexical information in $\text{KurLex}_{\text{tags}}$:

First, we developed a *guesser*, i.e., a tool that is able to guess the POS of a word that is unknown to the lexicon. Since MElt takes into account not only contextual and lexical but also string features of the word to be tagged (prefixes, suffixes, presence of a capital letter, and so on), we trained an instance of MElt on a specific training corpus which has been created by considering each lexical entry *(form,tag)* in $\text{KurLex}_{\text{tags}}$ as a sentence containing one word (*form*) tagged as *tag*. Of course, no external lexicon has been provided for training this special instance of MElt. Provided with an unknown word (e.g., a word that is not in KurLex and that is therefore not in the guesser's training corpus), this guesser outputs the word tagged with its guessed tag.

Second, we tokenised our corpus (excluding the small evaluation subcorpus) and projected $\text{KurLex}_{\text{tags}}$ on the tokenised corpus, i.e., we assigned to each word a disjunction of all the tags it is associated with in the lexicon. For words unknown to $\text{KurLex}_{\text{tags}}$, we used our guesser to obtain one tag. The result is an ambiguously tagged corpus.

Third, we designed various disambiguation methods to eliminate tagging ambiguities on words that have more than one tag in the lexicon:[18]

Bigram$_{\text{LR}}$ A left-to-right tag bigram model trained on the ambiguously tagged corpus.

Bigram$_{\text{RL}}$ A right-to-left tag bigram model trained on the ambiguously tagged corpus.

PrefHeuristics A model in which tags have been ranked manually on a global basis (e.g., being a particle is more likely than being a common noun, if both are possible) and always chosing the best

[18] We also tried a fourth method, that we called **Guesser**, which selects the tag that is best ranked by the guessing model. However, this method was far less accurate than the three other methods (approx. 72% on our small evaluation corpus, see below).

ranked tag among those that are known to the lexicon.
Each of these heuristics produced a disambiguated tagged corpus.

Fourth, we automatically merged the three resulting tagged corpora by tagging each sentence with tags that optimize the agreement with all models: for each token, we selected the tag chosen by a majority of models; if all were different, we selected the one from the corpus with the highest inter-model agreement. Finally, if the token being processed was the first one of the sentence (no history of agreement), we selected a token randomly. The result is a corpus of tagged sentences.

Finally, we trained the MElt tagger, using this corpus as a training corpus and $\text{KurLex}_{\text{tags}}$ as an external lexicon.

In order to evaluate the accuracy of this tagger, we manually annotated the small evaluation sub-corpus that contains 168 tokens (13 sentences). The precision of our MElt tagger on this corpus is 85.7%.[19,20] When used as a pre-annotation tool for the manual development of a POS-annotated corpus, such an accuracy level has been shown to lead to almost optimal improvements in both the speed and the quality of manual POS annotation (Fort & Sagot, 2010): for English, a POS tagger with 81.6% accuracy and a POS tagger with 96.4% accuracy lead to comparable annotation quality and almost identical annotation speed.

5 Conclusion and Future Work

We built a morphological lexicon and a POS tagger for Kurmanji Kurdish (KurLex) from only three types of data: lexical information, a non-formalised reference grammar and raw corpora. The resulting POS tagger should be used as a pre-annotation tool for developing a POS-annotated corpus, hence preparing the construction of a more accurate MElt-based POS tagger that relies on KurLex.

However, our methodology can be generalised to even less resourced languages, i.e. languages for which no lexical information is available. A methodology for building a morphological lexicon is described in (Walther & Sagot, 2010), with a partial application on Sorani Kurdish. Once the lexicon has been built, the above mentioned methodology is again fully applicable.

Moreover, our lexicon was developed within the Alexina framework (Sagot, 2010). A fair number of lexical resources are already being developed within this framework, such as the Le*fff* for French (Sagot, 2010) , the Le*ff*e for Spanish (Molinero *et al.*, 2009), the Le*ff*ga for Galician, SkLex for Slovak (Sagot, 2005), PolLex for Polish (Sagot, 2007), EnLex for English, PerLex for Persian (Sagot & Walther, 2010) and SoraLex for Sorani Kurdish (Walther & Sagot, 2010). This may also pave the way for future cross-language NLP applications.

Finally, since syntactic closeness has already proven useful for obtaining good results in machine translation (Mahsut *et al.*, 2001), we plan on trying and using the tools and resources created for Kurmanji to help the development of resources and tools for Sorani Kurdish, which is even more resource-scarce.

---

[19] This small evaluation corpus allowed us to evaluate each of the three above-mentioned model separately. The best model is PrefHeuristics, with a 87.5% accuracy. This accuracy is better than that of the final tagger. However, this difference is not statistically significant. Moreover, the PrefHeuristics method generates by nature annotations with systematic biases, which is not (or less) the case of the final tagger. The accuracy of $\text{Bigram}_{\text{LR}}$ (resp. $\text{Bigram}_{\text{RL}}$) is 79.2% (resp. 85.7%).

[20] Although our evaluation corpus is too small to draw precise conclusions, and despite the fact that Kurmanji Kurdish and English are quite different languages, one can compare our 85.7% accuracy on a 36-tag tagset with the accuracy figures reported in the recent literature on English. If provided with a lexicon that covers all tokens in the corpus (which is not the case in this work), (Ravi & Knight, 2009) report a 92% accuracy on a 45-tag tagset and a corpus larger than ours (approx. 78,000 tokens). With a smaller lexicon, these authors only provide figures on a small 17-tag tagset, resulting in a task much easier than ours. With a lexicon covering all tokens in the corpus, they reach a 96.8% accuracy, which drops to 88.0% if the lexicon only covers tokens appearing 3 or more times. Actually, in this setting, that is closer to ours, their results are lower than that obtained with the technique described in (Toutanova & Johnson, 2008) (89.7%).

References

Brill E. 1995. Unsupervised learning of disambiguation rules for part of speech tagging. In *Proceedings of the Third Workshop on Very Large Corpora*, p. 1–13, Cambridge, Massachussets, USA.

Creissels D. 2006. *Syntaxe générale – une introduction typologique 1 : catégories et constructions*. Paris, France: Hermès-Lavoisier.

Cucerzan S. & Yarowsky D. 2002. Bootstrapping a multilingual part-of-speech tagger in one person-day. In *COLING-02: proceedings of the 6th conference on Natural language learning*, p. 1–7, Morristown, NJ, USA: Association for Computational Linguistics.

Dandapat S., Sarkar S. & Basu A. 2007. Automatic part-of-speech tagging for Bengali: An approach for morphologically rich languages in a poor resource scenario. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, p. 221–224, Prague, Czech Republic: Association for Computational Linguistics.

Denis P. & Sagot B. 2009. Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art pos tagging with less human effort. In *Proceedings of PACLIC 2009*, Hong-Kong, China.

Dixon C. 1994. *Ergativity*. Cambridge Sudies in Linguistics. Cambridge, UK: Cambridge University Press.

Evans R. & Gazdar G. 1990. *The DATR Papers: February 1990*. Rapport interne CSRP 139, University of Sussex, Brighton, UK.

Fort K. & Sagot B. 2010. Influence of Pre-annotation on POS-tagged Corpus Development. In *Proceedings of the Fourth ACL Linguistic Annotation Workshop*, Uppsala, Sweden.

Francopoulo G., George M., Calzolari N., Monachini M., Bel N., Pet M. & Soria C. 2006. Lexical Markup Framework (LMF). In *Proceedings of LREC'06*, Genoa, Italy.

Gambäck B., Olsson F., Argaw A. A. & Asker L. 2009. Methods for Amharic part-of-speech tagging. In *AfLaT '09: Proceedings of the First Workshop on Language Technologies for African Languages*, p. 104–111, Morristown, NJ, USA: Association for Computational Linguistics.

Goldberg Y., Adler M. & Elhadad M. 2008. EM can find pretty good HMM POS-taggers (when given a good start). In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL'08)*, p. 746–754, Columbus, Ohio, USA.

Goldwater S. & Griffiths T. L. 2007. A fully Bayesian approach to unsupervised part-of-speech tagging. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'08)*, p. 744–751, Prague, Czech Republic.

Mahsut M., Ogawa Y., Sugino K. & Inagaki Y. 2001. Utilizing agglutinative features in japanese-uighur machine translation. In *Proceedings of MT Summit VIII*, p. 217–222, Santiago de Compostela, Spain.

Merialdo B. 1994. Tagging English text with a probabilistic model. *Computational Linguistics*, **20**(2), 155–72.

Molinero M. A., Sagot B. & Nicolas L. 2009. A morphological and syntactic wide-coverage lexicon for Spanish: The Le*ff*e. In *Proceedings of the 7th conference on Recent Advances in Natural Language Processing (RANLP 2009)*, Borovets, Bulgaria.

Ravi S. & Knight K. 2009. Minimized models for unsupervised part-of-speech tagging. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP'09)*, p. 504–512, Singapore.

Sagot B. 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*, p. 156–163, Karlovy Vary, Czech Republic.

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

Sagot B. 2010. The Le*fff*, a freely available, accurate and large-coverage lexicon for French. In *Proceedings of LREC'10*, Valetta, Malta.

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

Sánchez-Martínez F., Pérez-Ortiz J. A. & Forcada M. L. 2008. Using target-language information to train part-of-speech taggers for machine translation. *Machine Translation*, **22**(1-2), 29–66.

Skoumalová H. 1997. A Czech morphological lexicon. *Proceedings of the Third Meeting of the ACL Special Interest Group in Computational Phonology*.

Smith N. & Eisner J. 2005. Contrastive estimation: Training log-linear models on unlabeled data. In *Proceedings of the 43th Annual Meeting of the Association for Computational Linguistics (ACL'05)*, p. 354–362, nn Arbor, Michigan, USA.

Thackston W. M. 2006. Kurmandji Kurdish: A reference grammar with selected readings. http://www.fas.harvard.edu/ iranian/Kurmanji/kurmanji_1_grammar.pdf.

Toutanova K. & Johnson M. 2008. A Bayesian LDA-based model for semi-supervised part-of-speech tagging. In *Proceedings of the 21st International Conference on Advances in Neural Information Processing Systems (NIPS)*, p. 1521–1528, Vancouver, British Columbia, Canada.

Walther G. & Sagot B. 2010. Developing a Large-Scale Lexicon for a Less-Resourced Language: General Methodology and Preliminary Experiments on Sorani Kurdish. In *Proceedings of the 7th SaLTMiL Workshop on Creation and use of basic lexical resources for less-resourced languages (LREC 2010 Workshop)*, Valetta, Malta.