Developing a lexicon of word families for closely-related languages

Núria Gala

▶ **To cite this version:**

Núria Gala. Developing a lexicon of word families for closely-related languages. ESSLLI International Workshop on Lexical Ressources, Aug 2011, Ljubljana, Slovenia. pp.41-47. hal-01778945

HAL Id: hal-01778945
https://hal.science/hal-01778945

Submitted on 26 Apr 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Developing a lexicon of word families for closely-related languages

Nuria Gala

LIF-CNRS UMR 6166
163 Av. de Luminy case 901 F-13288 Marseille cedex 9
nuria.gala@lif.univ-mrs.fr

Abstract

Lexical resources are of interest in linguistic research and its applications. However, building and enriching them is very time consuming and expensive. In specific fields such as morphology, unsupervised and (semi-)supervised approaches consisting in automatically discovering word structure have gained in popularity in the last few years. While encouraging results have been obtained for a large variety of languages, few resources are currently available. In this paper, we describe a morphological lexicon under development for Romance languages. It is based on an initial seed set of manually identified 2,004 word families in French. Our goal is to map these families on related languages in order to obtain a resource based on family clusters, capable to provide morphological and semantic information on each family crosslingually. Such a resource will be of help in contrastive linguistics and in different NLP and human applications, such as crosslingual information retrieval and interlingual language learning.

1. Introduction

A variety of multilingual lexical resources have been developed by different civilizations ever since the birth of writing, as a result of practical needs (learning, archiving, transmitting linguistic and other kind of knowledge, etc.). The shape and the contents of these resources have evolved significantly over time. Technical revolutions such as printing and computerization have had a profound influence on the way to develop lexicons. From linear presentations of word lists to lexical networks, multilingual lexical resources present interlingual correspondences and often very specific linguistic information.

Obviously, manually building and enriching such resources is very time consuming and expensive. In recent years, collaborative and automatic approaches have emerged as a plausible alternative to build resources in a large-scale perspective thus limiting the time of development. Collaborative multilingual resources such as *Papillon* (Boitet et al., 2002) are based on the principle of sharing contributions, that is, anyone collaborates to enrich the database according to his/her possibilities. While the underlying philosophy is interesting, the results can easily be disappointing, as enriching a resource is a tedious task and in practice few people accept. Hence, it is hard to get the expected volume of contribution[1].

In order to address both shortcomings (manual and contributive), automatic approaches have gained in popularity in NLP, especially when it comes to collecting specific linguistic information. In morphology, different methods exist to automatically acquire information about the internal structure of words (Lavalle and Langlais, 2010): probabilistic approaches which regroup words into paradigms by removing common affixes (*Paramor* (Monson et al., 2007)) or community-based detection in networks (*MorphoNet* (Bernhard, 2010)), unsupervised learning of word structure by decomposition (*Linguistica* (Goldsmith, 2001), *Morfessor* (Creutz and Lagus, 2005)), supervised or semi-supervised methods using formal analogies to identify stems and morphological information (Lepage, 1998), (Hathout, 2008), (Lavalle and Langlais, 2010).

These methods may differ with respect to the kind of result they obtain: word segments, complete morphemic analysis, morphological links between words, etc. Furthermore, as raw text is used for knowledge acquisition, most systems do not make a difference between inflectional or derivational morphology.

The work presented in this paper aims at building crosslinguistic morpho-phonological families. A morpho-phonological family groups lexical units sharing morphological[2] and semantic features. Such a family is usually built around a common stem. Hence, the stem *terre* 'earth', will induce the family made of lexical units such as *terrasse* 'terrace', *terrestre* 'terrestrial', *terrien* 'landowner', etc. All the words in this family share the following semantic components: 'surface', 'ground', 'area', etc. Having translated the stem in closely-related languages and using multilingual corpora and lexica, we will build the corresponding families and compare their organization across languages. Our aim is thus to create a resource presenting word families among closely-related languages and to check whether they can be mapped on each other. The linguistic description provided is strictly synchronical and concerns both derivational morphology (stems and affixes) and morphosemantic links (semantic components within a word family).

This paper is structured as follows. In the next section we provide an overview concerning some existing mono- and multilingual resources by focusing on those containing a morphological description. Section 3 describes first experiments to map our initial resource for French to other Romance languages. We conclude the paper by discussing the achieved results and present some ideas concerning future work.

[1] For a discussion, see (Cristea et al., 2008).

[2] Phonological alternations are possible, i.e. *fleur/flor-* in *fleur* 'flower' and *floraison* 'flowering', *croc/croch-* in *croc* 'hook' and *crochet* 'little hook'.

2. Morphological resources: an overview

Although a significant number of existing NLP lexicons present primarily syntactic or semantic information — subcategorization (Briscoe and Carrol, 1997), concepts as in *WordNet* (Fellbaum, 1998), etc. — an increasing interest in morphology has led over the past few years to the development of morphological lexica. Such resources present a fine-grained and explicit description of the morphological organization of the lexicon. The resources are mainly monolingual, though some multilingual examples can be mentioned.

2.1. Monolingual lexica

For morphology rich languages such as Romance or Slavic languages, monolingual lexica may display morphotactics (ordering of morphemes, derivational morphology) or morphosyntactic information (word forms associated to: a lemma, a part-of-speech tag, inflectional categories, subcategorization patterns, etc.).

The *Digital Dictionary of Catalan Derivational Affixes (DSVC)* (Bernal and DeCesaris, 2008) illustrates a derivational morphology lexicon. It has been created manually and is of limited coverage: about one hundred verbs. *Lefff* (Clement et al., 2004) is an example of morphosyntactic lexicon for French verbs (about 5,000 entries). It has been built automatically by extracting information from large raw corpora and other existing resources.

As for Slavic languages, *Unimorph*[3] is a derivational morphology database with 92,970 Russian words. There is also a morphosyntactic lexicon for Polish (Sagot, 2007) which has been created using the same formalism as the one in *Lefff* through automatic lexical acquisition from corpora. Such a formalism has also been used for other European languages (e.g. Spanish), as well as for less resourced languages such as Kurdish and Persian (Walther and Sagot, 2010).

2.2. Multilingual lexica

Multilingual resources provide the basis for translation, that is, the mapping from one language to the other (Calzolari et al., 1999). Yet this does not always hold for all multilingual morphological resources.

A leading example is CELEX (Baayen et al., 1995), a manually-tagged morphological database for English, Dutch and German. For each language, words are analyzed morphologically and the processes of derivation are made explicit (e.g. 'concern'[V], 'unconcern' ((un)[N—.N], ((concern)[V])[N])[N]). Unfortunately, the morphological information is not explicit crosslinguistically, that is, CELEX is a database for three languages independent one from another.

MuleXFoR[4] (Cartoni and Lefer, 2010) is a morphological database aiming to present word-formation processes in a multilingual environment. Word formation is presented as a set of multilingual rules available by affixes, rules and constructed words (e.g. by the rule *'above (n < a)'*, the following affixes are displayed: *sopre, sovra, super* (Italian), *sur, supra* (French), *supra* (English), along with some words containing such affixes in each language). Word formation processes are thus represented in a multilingual context. Although morphological knowledge was partly automatically acquired from corpora, the coverage of MuleXFoR is limited to one hundred prefixes.

Finally, unsupervised learning of morphologically related words in various languages (English, German, Turkish, Finnish and Arabic) has been the main goal of systems participating to Morpho Challenge 2009[5], e.g. *Morfessor* (Creutz and Lagus, 2005), *Rali-Cof* (Lavalle and Langlais, 2010), *MorphoNet* (Bernhard, 2010), etc. While such competition allows the comparison of different statistical machine learning techniques (in terms of precision and recall), the challenge does not yield any avalaible morphologically annotated resource.

2.3. Remarks

Two general observations can be made at this point: first, very few available resources present morphological links crosslinguistically, and if they do, their coverage is limited; second, morphological processes described by the existing resources mainly focus on word-formation (word construction) conveyed by affixes. To our knowledge, word families — although described in the literature (Bybee, 1985) — have been brought to the forefront only in psycholinguistics to show their impact in lexical decision tasks (Schreuder and Baayen, 1997).

3. Mapping from French to other Romance languages

Considering that closely-related languages have a common origin, morphological regularities may be conveyed by means of similar constructions. Our aim is thus to use a manually built morphological lexicon (*Polymots*[6] (Gala et al., 2010), with 2,004 stems and nearly 20,000 derived words for French) and map it to other Romance languages.

3.1. Word families: definition and properties

We consider a family (cluster or paradigm) to be a set of lexical units sharing a formal and a semantic component. Similar words in a lexical cluster share:

- a stem (e.g. *human* in *human*, *humanism, humanist, humanitarian, humanity, humanize, dehumanize*, etc.);
- semantic continuity (all the words in the previous serie are related to the notion of 'bipedal primate mammal').

While in some families there is a continuity of meaning (words sharing a significant number of semantic features, e.g. the *human* family), in others meaning is distributed, i.e. a single and precise meaning is impossible to seize among the lexical units of the cluster, as the words have evolved and the semantic components are widely dispersed.

[3] http://courses.washington.edu/unimorph/

[4] http://www.issco.unige.ch/en/staff/bruno/mulexfor

[5] http://research.ics.tkk.fi/events/morphochallenge2009/

[6] http://polymots.lif.univ-mrs.fr

In such cases, the semantic features of the common stem are to be found among the words in the family (e.g. French *val-* 'glen' includes features such as *geographic area* and *going downhill* and at least one of these notions is to be found in *vallée* 'valley' and *avaler* 'swallow'). Semantic components have been automatically extracted from structured corpora (Gala and Rey, 2009) and are currently being refined with a new machine-readable lexicographic resource (the *Trésor de la Langue Française informatisé*).
Similarly, the size of the families may vary significantly depending on the stem, going from no derivation at all (e.g. French *agrume* 'citrus fruit', *paupière* 'eyelid', etc.) to more than eighty derived words (e.g. *port-* in *apport, emporter, exporter, porable, etc.*). The larger the family the more significant the semantic dispersion; however, the higher the number of analogous word-forms across languages.

3.2. From French to other Romance languages

The French lexicon that we have used to map to the other Romance languages contains 2,004 stems (i.e. 2,004 families). The stems are of two kinds: 87 % (1,741) are lexemes (e.g. *terre* 'ground', *bras* 'arm', etc.); the remaining 13 % (263) are word-forms which do not exist anymore as single tokens (stems in italics, e.g. *bast*ille, *bast*ion, etc.; ap*port*, ex*port*er, etc.; con*verg*ence, di*verg*ent, etc.).
From the initial 1,741 stems, we have conducted a preliminary experience by using a subset of 30 stems (see table 1). To obtain these 30 words, we have selected the most frequent ones from the Greenberg's list, i.e. those having a frequency $f > 0,1$ % on the BNC (70 out of 100). Once these 70 words have been manually translated into French, we have kept those having a frequency $f > 0,01$% in the VocaRef corpora[7] (Table 1). For each word in this list we have automatically acquired their lexical clusters (families) using raw corpora, POS tagged and lemmatized corpora.

grand, dire, voir, homme, venir, donner, savoir, petit, bon, nouveau, personne, femme, entendre, tête, nom, nuit, eau, long, sein, coeur, pierre, humain, mourir, tuer, langue, feu, chemin, bras, sang, oeil
big, say, see, man, come, give, know, small, good, new, person, woman, hear, head, name, night, water, long, breast, heart, stone, human, die, kill, tongue, fire, path, arm, blood, eye

Table 1: List of 30 words from Greenberg's list with $f >$ 0,1 % in the BNC and $f >$ 0,01 % in VocaRef corpora.

3.2.1. Semi-supervised learning using raw corpora

After having translated the 30 seed words in Spanish, Catalan, Italian, Corsican and Portuguese, we extracted all the words from different raw corpora[8] containing every single stem (e.g. in Catalan, *branca*da, *branca*l, *branca*tge, etc. contain the stem *branca* 'branch'). We have refined such first loop with three variants. First, if the stem ends in a vowel, we have retrieved all the forms containing the stem minus the final vowel. We did this in order to address the problem of vocal alternations (e.g. Italian *nome* 'noun, name' / *nominare* 'nominate'). Second, if the stem ends in a voiceless velar plosive (/k/) or a voiceless alveolar fricative (/s/) we had a look at all possible graphical variants: e.g. for the latter, Portuguese and Catalan ç / c (*cabeça / cabecear* 'head / to head', *braç / bracet* 'arm / little arm'). Finally, considered alternations for diphthongs in Spanish: /e/ with /je/ (*pedrería / piedra* 'jewels / stone'), /o/ with /we/ (*novedad / nuevo* 'novelty / new').
The words obtained were then been manually validated via a monolingual dictionary for each one of the respective languages. This allowed us also (1) to capture words absent from the corpora and (2) to eliminate candidates wrongly retrieved because we had used only formal analogies without taking into account any semantic information (e.g. in Corsican, the stem *testa* 'head' yields *testa*tu 'stubborn' and in*testa*tura 'header', while cun*testa* 'answer' and *testa*mentu 'will' do not show up in the expected 'head' cluster, eventhough they present the same graphical form). Table 2 shows several members crosslingually gathered for the same family.

FR	oeil	oeillade, oeillard, oeillère, oeillet...
CA	ull	ullada, ullera, ullerat, ullerer, ullerol...
CO	ochju	malochju, ochjuculà, ochjutu...
PT	olho	olhar, olhadinha, olheiras, olhudo...
ES	ojo	ojera, ojear, ojal, ojoso, ojuelo...
IT	occhio	occhialàio, occhiàle, occhialìno...

Table 2: Exemples of 'eye' family.

This first experience reveals that using the stem and some morpho-phonological variants allows us to gather a significant number of candidates belonging to the same family. Not surprisingly, the longer the stem, the higher the accuracy (see 4.1).

3.2.2. Semi-supervised learning based on annotated corpora

A second experience has been carried out to map the initial French stems to other closely-related languages. This time we wanted to scale up to a higher number of families using the same heuristics as the one used in the previous test (stems and their morpho-phonological variants). We also wanted to restrict the mapping to two languages (Catalan and Spanish).
The underlying hypothesis for this second experience is the idea that using annotated corpora would increase the accuracy of the results, mainly because of the absence of inflection. This being so, we used a POS tagged and lemmatized corpora extracted from Wikipedia (258,315 lemmas for Catalan, 387,003 for Spanish)[9]. The corpora have been annotated with lemma and part of speech information using the open source library Freeling[10] 2.1.

[7] 234 millions of words from French newspapers *Libération* and *Le Monde*, 1995-1999.

[8] We have extracted corpora from the Web, mainly Wikipedia and newspaper sites.

[9] http://www.lsi.upc.edu/~nlp/wikicorpus/

[10] http://nlp.lsi.upc.edu/freeling/

We used bilingual corpora to automatically extract the translations of our initial 1,741 stems which are lexemes. The corpora used[11] (7,523 entries FR-CA and 25,616 entries ES-CA) allowed us to extract 30 % of the expected trilingual equivalences, that is 473 stems out of the initial 1,741. From such trilingual set of stem equivalences, we have gathered word forms in the Wikipedia corpora containing each stem and its variants for each of the two languages. At the end of the experience, we have obtained 190 families for Spanish (40,2 %) and 77 for Catalan (16,3 %).

4. Preliminary results

4.1. Raw corpora and stem lengths

The results of hand-validating the data obtained for five languages from raw corpora shed light on significant differences related to the length of words. As we have taken a list of very frequent words, the global average length for all languages is 5 characters as shown in Table 3.

Language	Average	$>= 5$	< 5
CA	4,3	43 %	57 %
CO	4,7	57 %	43 %
FR	4,9	53 %	46 %
PT	5,1	67 %	33 %
ES	5,2	80 %	20 %
IT	5,6	80 %	20 %

Table 3: Word lengths.

Taking into account such a threshold (i.e. 5 characters), precision is about 85 % for stems of length greater or equal to 5 and about 15 % for stems with less than 5 characters. The shorter the stem, the higher the number of word-forms collected, only few being members of the expected family. Furthermore, the shorter the stem, the higher the possibility of homonymy (e.g. in Catalan, *nou* 'new / nine / walnut'), hence the higher the probability to collect word-forms valid from a formal point of view, but unacceptable semantically in a given paradigm (e.g. *noucentista* 'related to the beginnings of years 1900', hence related to 'nine' but not to 'new'). It is also noticeable that in raw corpora inflected and compounded word-forms, as well as misspelled words (e.g. **dinousaure* 'dinosaur'), contribute to decrease the precision rate for all languages (we aimed at collecting only well-formed derived lexical units). As for recall, the data has been manually evaluated by comparing the words obtained with a list of entries present in a monolingual dictionary for each language. As we have used relatively small corpora (100,000 to 300,000 words) global recall is about 50 %, again with significant differences among languages and families.

4.2. Bilingual corpora and analogies among languages

The availability and the size of the resources is crucial for semi-supervised acquisition of information. In our experience, only 30 % of the stems have a correspondence in the three languages. Bigger bilingual corpora is thus necessary to scale up automatically to our initial 2,004 stems as well as to map from French to other languages. At this stage, the comparison of the 473 stems among French, Spanish and Catalan, already gives us some insights concerning the relative linguistic distance of these three languages (cognates, see table 4). Our aim is to consider cognates among families and not only among individual words.

Analogies		example FR CA ES
FR CA ES	69,98 %	*ouvert obert abierto* (open)
FR CA -	5,29 %	*pleur plor llanto* (cry)
- CA ES	16,91 %	*besoin necessitat necesidad* (need)
FR - ES	1,48 %	*corne banya cuerno* (horn)
- - -	6,34 %	*creux buit hueco* (hollow)
	100 %	

Table 4: Lexical closeness among FR CA ES.

About 70 % of the stems are analogous in the three languages, i.e. lexical items share the same form; Catalan and Spanish are closely-related in 86,89 %, Catalan and French in 75,27 % and French and Spanish in 71,46 %.

4.3. Lemmatized corpora and family clusters

The use of a large coverage corpus has enabled us to obtain family clusters very quickly: we have gathered 5,999 word-forms being part of 190 families in Spanish and 1,561 word-forms for 77 families in Catalan. We have conducted an evaluation on 40 stems (40 different families) with 618 words in Spanish and 428 words in Catalan. The average precision is 62,71 % for Spanish and 64,42 % for Catalan, but the results are very heterogeneous among the families. Yet for some families precision is very high (in some cases, 100 %, i.e. all the acquired words belong to the family, see table 5). However, in other families, precision is very low (the word-forms obtained do not belong to the cluster) mainly for reasons of homography (e.g. in Catalan, the string *tendre* 'tender' can be found at the end of a significant number of verbs *abstendre, desentendre, entendre, estendre, etc.*). Some drawbacks come also from the corpora itself: words in other languages, misspellings and errors in tokenization and lemmatization. With the heuristics employed (stems and morpho-phonological variants), we also capture all the existing compounds for a given stem. We are thus considering whether to include them into the resource or to limit it to derivation strictly. Recall is under evaluation using monolingual dictionaries for both languages.

abrigo	abrigo, abrigado, abrigador, abrigamiento, abrigar, desabrigar, desabrigado, desabrigo
aceituna	aceituna, aceitunado, aceitunero, aceitunillo
chocolate	chocolate, achocolatado, chocolatada, chocolatería, chocolatero, chocolatina

Table 5: Family clusters for Spanish.

[11] http://sourceforge.net/projects/apertium

5. Conclusion

In this paper, we have presented some initial steps in developing a lexical resource for word families across closely-related languages. The lexicon is constructed from an initial set of stems, identified and manually validated for French. The approach relies on automatically acquiring information from corpora and it reveals that using partially annotated corpora (lemmatized corpora) leads to better results provided that morphophonological properties of each language (e.g. diphthongs in Spanish, consonant and vocalic alternations) are taken into account. We have thus automatically acquired lexical clusters with equivalences in closely-related languages. Word families are connected in order to allow crosslingual access of lexical items via morphological and/or semantic criteria.

Such a lexicon is under development at the time of writing this paper: scaling up to the initial 2,004 stems will be carried out soon for Catalan and Spanish (with bigger annotated corpora and lexica available).

As for future work, we plan to extend the resource to the remaining Romance languages[12]. Furthermore, automatic acquisition of morphological information on analogical series of words (words containing the same affixes, (Hathout, 2008)) is also foreseen shortly. The resulting resource will be freely available and we hope it will be of help for many multilingual human and NLP applications.

6. Aknowledgements

The author would like to thank M. Zock as well as the anonymous reviewers for their valuable comments and suggestions to improve the quality of the paper.

7. References

R. H. Baayen, R. Piepenbrock, and L. Gulikers. 1995. The celex lexical database release 2. CD-ROM. Linguistic Data Consortium, Univ. of Pensilvania, USA.

E. Bernal and J. DeCesaris. 2008. A digital dictionary of catalan derivational affixes. In *Proceedings of Euralex 2008*, Barcelone, Spain.

D. Bernhard. 2010. Morphonet: Exploring the use of community structure for unsupervised morpheme analysis. In *Multilingual Information Access Evaluation. 10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009. Revised Selected Papers.*, volume 1, pages x–x. Springer.

C. Boitet, M. Mangeot, and G. Serasset. 2002. The papillon project: cooperatively building a multilingual lexical data-base to derive open source dictionaries and lexicons. In N. Ide G. Wilcock and L., editors, *Proceedings of on Natural Language Processing and XML, COLING Workshop*, pages 9–15, Taipei, Taiwan.

T. Briscoe and J. Carrol. 1997. Automatic extraction of subcategorization from corpora. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, Whashington, DC.

[12]Italian, Corsican, Portuguese, Galician and Romanian, possibly Sardinian — the oldest Romance language — provided that enough data is available, i.e. machine-readable annotated corpora and monolingual dictionaries).

J. L. Bybee. 1985. *Morphology. A study of the relation between meaning and form. Typological studies in Language*. Benjamins, Amsterdam.

N. Calzolari, K. Choukri, C. Fellbaum, E. Hovy, and N. Ide. 1999. Multilingual ressources. *Multilingual Information Management : current levels and future abilities. Report commissioned by the US National Science Foundation and the European Commission's Language Engineering Office*.

B. Cartoni and M. A. Lefer. 2010. The mulexfor database: Representing word-formation processes in a multilingual lexicographic environment. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valetta, Malta.

L. Clement, B. Sagot, and B. Lang. 2004. Morphology based automatic acquisition of large-coverga lexica. In *Proceedings of the Fourth conference on International Language Resources and Evaluation (LREC'04)*, Lisbonne, Portugal.

M. Creutz and K. Lagus. 2005. Unsupervised morpheme segmentation and morphology induction from text corpora using morfessor 1.0, publications in computer and information science. Technical Report A81, Helsinki University of Technology.

D. Cristea, C. Forascu, M. Raschip, and M. Zock. 2008. How to evaluate and raise the quality in a collaborative lexicographic approach. In *Proceedings of the Sixth conference on International Language Resources and Evaluation (LREC'08)*, Marrakech.

C. Fellbaum. 1998. Wordnet: an electronic lexical database. Technical report, MIT Press, Cambridge, MA.

N. Gala and V. Rey. 2009. Acquiring semantics from structured corpora to enrich an existing lexicon. In *Electronic lexicography in the 21st century: new applications for new users (eLEX-2009)*, Louvain-la-Neuve, Belgium.

N. Gala, V. Rey, and M. Zock. 2010. A tool for linking stems and conceptual fragments to enhance word access. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valetta, Malta.

J. Goldsmith. 2001. Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, 27:153–198.

N. Hathout. 2008. Acquisition of the morphological structure of the lexicon based on lexical similarity and formal analogy. In *3rd Textgraphs workshop*, pages 1–8, Manchester, UK.

J. F. Lavalle and Ph. Langlais. 2010. Apprentissage non supervisé de la morphologie d'une langue par généralisation de relations analogiques. In *Proc. Traitement Automatique des Langues Naturelles (TALN-10)*, pages x–x, Montreal.

Y. Lepage. 1998. Solving analogies on words: an algorithm. In *Proceedings of the 17th international conference on Computational linguistics (COLING)*, pages 728–735, Montreal.

C. Monson, J. Carbonell, A. Lavie, and L. Levin. 2007. Paramor: minimally supervised induction of paradigm structure and morphological analysis. In *Proceedings of*

9th SIGMORPHON Workshop, pages 117–125, Prague, Czech Republic.

B. Sagot. 2007. Building a morphosyntactic lexicon and a pre-syntactic processing chain for polish. In *Proceedings of the 3rd Language and Technology Conference*, pages 42–427, Pozna, Poland.

R. Schreuder and R. H. Baayen. 1997. How complex simple words can be. *Journal of Memory and Language*, 53:496–512.

G. Walther and B. Sagot. 2010. Developing a large-scale lexicon for a less-resourced language : General methodology and preliminary experiments on sorani kurdish. In *Proceedings of the 7th SaLT-MiL Workshop on Creation and use of basic lexical resources for less-resourced languages (LREC 2010 Workshop)*, Valetta, Malta.