

Experiences with Developing Language Processing Tools and Corpora for Amharic

Björn GAMBÄCK^{1,2}, Lars ASKER^{3*}

¹*SICS, Swedish Institute of Computer Science AB, Box 1263, 164 29 Kista, Sweden*

Tel: +46 8 633 15 35, Fax: +46 8 751 72 30, Email: gamback@sics.se

²*Computer & Information Science, Norwegian Univ. Sci & Tech, 7491 Trondheim, Norway*

Tel: +47 735 933 54, Fax: +47 735 944 66, Email: gamback@idi.ntnu.no

³*Computer & System Sciences, Stockholm University, Forum 100, 164 40 Kista, Sweden*

Tel: +46 8 674 70 02, Fax: +46 8 703 90 25, Email: asker@dsv.su.se

Abstract: A major bottleneck for promoting use of computers and the Internet is that many languages lack access to basic tools that would make it possible for people to access ICT in their own language. The paper describes the development a set of such resources for the processing of Amharic, the working language of the Ethiopian government. The primary goal was to investigate techniques and methods that can be used to efficiently create computational linguistic resources for new languages based on existing tools and resources. The resources created consist of linguistically annotated text collections and tools for word-level analysis of Amharic.

Keywords: Part-of-Speech Tagging, Text Categorization, Corpora, Amharic.

1. Introduction

There is a need for people all over the World to be able to use their own language when using computers or accessing information on the Internet. This requires the existence of a variety of applications including local language spell-checkers, word processors, machine translation systems, search engines, etc. At the same time, the amount of work required to develop all aspects of natural language processing for a new language is huge.

Even though the last years have seen an increasing trend in applications of language processing methods to languages other than English, most of the work is still done on very few and mainly European and East-Asian languages. Many languages, especially on the African continent, are “under-resourced” in that they have very few computational linguistic tools or corpora (such as lexica, tree-banks, part-of-speech taggers, parsers, etc.) available. This has proven to be a bottleneck for promoting the use of computers and the Internet in a language, with the basic problem being related to the way that the richer in general tend to get richer and the poorer get poorer, here manifesting itself in that it is more difficult to develop new linguistic resources without having access to already existing ones.

Until a critical mass of language processing resources have been made publicly available, it is not likely that natural language processing for any language can “take off” and reach a level higher than that of academic prototype systems. Making such resources publicly available will further boost activities in the area by allowing researchers to benefit from the work of others and thereby to build more advanced tools without having to invent the wheel over and over again. We believe this to be the best chance for any language with limited computational linguistic resources to break the vicious circle of always having to start from scratch with research in the area.

* This work has been partially funded by Sida, the Swedish International Development Cooperation Agency through SPIDER, the Swedish Programme for ICT in Developing Regions (www.spidercenter.org).

With this in mind, the primary goal of the work described in this paper has been to investigate how well existing linguistic knowledge can be transferred between languages and to develop resources, tools and techniques that can support such knowledge transfer. We have worked with Amharic, the official working language of the federal government of the Federal Democratic Republic of Ethiopia.

Amharic is spoken by about 30 million people as a first or second language, making it the second most spoken Semitic language in the world (after Arabic), probably the second largest language in Ethiopia (after Oromo, a Cushitic language), and possibly one of the five largest languages on the African continent. The actual size of the Amharic speaking population must be based on estimates: Hudson [20] analysed the Ethiopian census from 1994 and indicated that more than 40% of the population then understood Amharic, while the current size of the Ethiopian population is about 80 million. (85.2 million according to the CIA [11]; 76.9 million according to Ethiopian parliament projections in December 2008 based on the preliminary reports from the census of May 2007).

Following the Constitution drafted in 1993, Ethiopia is divided into nine fairly independent regions, each with its own nationality language. However, Amharic is the language for country-wide communication and was also for a long period the principal literal language and medium of instruction in primary and secondary schools of the country, while higher education is carried out in English.

In spite of the relatively large number of speakers, Amharic is still a language for which very few computational linguistic resources have been developed, and very little has been done in terms of making useful higher level Internet or computer-based applications available to those who only speak Amharic. It is generally believed that applications such as information retrieval, text classification, or document filtering could benefit from the existence and availability of basic tools such as stemmers, morphological analysers or part-of-speech taggers. However, since so few language processing resources for Amharic have been available, very little is known about their effect on retrieval or classification performance for this language. (See [24] for a more general discussion about available tools for Semitic languages, even though mainly concentrating on Arabic, Hebrew and Maltese. In that paper, Wintner also gives several arguments for the need of developing language resources for these languages, and the problems faced when doing that.)

The rest of this paper is outlined as follows. First the Amharic language is described in some detail in Section 2. The sections thereafter discuss the activities and results from the project, first the collection, building and annotation of the necessary corpora in Section 3, then in Sections 4-6 the main tools created in the project: three part-of-speech taggers, a shallow morphological analyser (a stemmer) for Amharic, and text classifiers. Section 7 discusses applications and further refinement of the resources, and draws some conclusions.

2. Amharic

Written Amharic (together with the closely related Tigrinya language) uses a unique script which has originated from the Ge'ez alphabet (the liturgical language of the Ethiopian Orthodox Church). Written Ge'ez can be traced back to at least the 4th century A.D. Unlike Arabic or Hebrew, the language is written from left to right. The first versions of the writing system included consonants only, while the basic characters in later versions mainly represent consonant-vowel (CV) phoneme pairs. The script also has a unique set of punctuation marks and digits, and some special characters for labialised consonants.

In the modern Ethiopic script each basic syllable pattern comes in seven different forms (called *orders*), reflecting the seven vowel sounds. The first order is the basic form; the other orders are derived from it by more or less regular modifications indicating the different vowels. There are 33 basic forms, giving 7×33 syllable patterns called *fidEls*. Two of the base forms represent vowels in isolation, but the rest are for consonants (or semi-

vowels classed as consonants) and thus correspond to CV pairs, with the first order being the base symbol with no explicit vowel indicator. In total, there are 275 *fidEls*.

Like many other Semitic languages, Amharic has a rich verb morphology which is based on triconsonantal roots with vowel variants describing modifications to, or supplementary detail and variants of the root form. The Amharic writing system uses multitudes of ways to denote compound words and there is no agreed upon spelling standard for compounds. As a result of this – and of the size of the country leading to vast dialectal dispersion – lexical variation and homophony is very common.

In addition, not all the letters of the Amharic script are strictly necessary for the pronunciation patterns of the spoken language; some were simply inherited from Ge'ez without having any semantic or phonetic distinction in modern Amharic. There are many cases where numerous symbols are used to denote a single phoneme, as well as words that have extremely different orthographic form and slightly distinct phonetics, but with the same meaning. So are, for example, most labialised consonants basically redundant, and there are actually only 39 context-independent phonemes (monophones): of the 275 symbols of the script, only about 233 remain if the redundant ones are removed.

3. Corpora and Annotation

The activities within the project described in the present paper can be broken down into the following parts: corpora collection and (manual) tagging, (automatic) part-of-speech tagging, morphological analysis, and further refinement and application of the resources. These different activities are discussed in turn in the following sections, starting with the corpus building.

3.1. Unannotated Corpora

Several Amharic corpora have been collected [6]. The largest one consisting of approximately 3.5 million words of Amharic news text (in transliterated form) from Ethiopian News Headlines (www.ethiozena.net).

Another corpus was collected from Walta Information Center (www.waltainfo.com) and consists of 8715 Amharic news articles from the years 2001–2004. A subset of these articles has been manually annotated with appropriate part-of-speech tags by human annotators (see Section 3.3 below).

Two Amharic–English parallel corpora consisting of government policy files have also been collected from the web pages of the Ethiopian Ministry of Information (www.moinfo.gov.et).

3.2. Tag sets

Three different suitable and appropriate tag sets for Amharic have been used in the corpus annotation. The first two sets were developed by linguists and faculty at AAU, Addis Ababa University [12]. The basic tag set consists of 10 classes: Noun, Pronoun, Verb, Adjective, Preposition, Conjunction, Adverb, Numeral, Interjection, and Punctuation plus one extra tag for problematic (unclassified) words. Thus the first tagset has 11 tags.

Some of the 10 basic classes can be further divided into subclasses, creating a second set with a total of 30 tags. The main differences between these two tag sets pertain to the treatment of prepositions and conjunctions: in the larger set there are specific classes for pronouns attached with preposition, conjunction, and both proclitic preposition and enclitic conjunction (similar classes occur for nouns, verbs, adjectives, and numerals). In addition, numerals are divided into cardinals and ordinals, verbal nouns are separated from other nouns, while auxiliaries and relative verbs are distinguished from other verbs.

The third tag set has 10 tags and is the one introduced by Sisay [14], where the tag set was used in part-of-speech tagging experiments based on Conditional Random Fields. These classes include one for Residual which was assumed to be equivalent to unclassified group in the other tag sets. In addition, conjunctions and prepositions are not separated, but jointly tagged as adpositions, and pronouns are taken to belong to the noun class. The other classes were assumed to be the same ones as in the basic tag set above, except that the basic tag set from AAU groups all verbs together, while Sisay kept Auxiliary as its own class.

3.3. *Annotated Corpora*

A corpus of 1065 Amharic news articles (originally 210,000 words in the Ethiopic script) has been manually annotated with appropriate lexical attributes (part-of-speech) with the help of annotators from the Ethiopian Languages Research Center at Addis Ababa University, using the full (30-tag) tag set described above [12]. The tagged corpus has been made freely available on the Internet in XML-format in both Ethiopic script and in a transliterated form. It can be downloaded from nlp.amharic.org. The original Amharic text was transliterated into SERA (System for Ethiopic Representation in ASCII) using a file conversion utility called g2 which is available in the LibEth package (g2 was made available to us by Daniel Yacob of the Ge'ez Frontier Foundation (www.ethiopic.org)).

Unfortunately, the manually annotated corpus contains quite a few errors and tagging inconsistencies: nine persons participated in the manual tagging, writing the tags with pen on hard copies, which were given to typists for insertion into the electronic version of the corpus – a procedure obviously introducing several possible error sources.

In a second round [17], the corpus has thus been “cleaned” semi-automatically: many non-tagged items needed to be tagged (in the first version of the corpus there are, for example, many cases where the headlines of the news texts had been tagged as one item, end-of-sentence punctuation), while some double tags had to be removed and multiword units with single tags were merged to one word. Furthermore, several tagging errors and misspellings were corrected, while tagging inconsistencies were levelled out.

After transcription and corrections, the size of the corpus was 200,863 words (in the SERA form). This corrected and transcribed corpus was automatically mapped to the two smaller tag sets (with 11 and 10 tags, respectively) described in the previous section.

In addition, the same corpus of 1065 Amharic news articles has been manually annotated with appropriate class labels to be used in experiments concerning automatic text classification, dividing the texts into ten different news categories [5, 4].

4. Part-of-Speech Tagging

Part-of-speech tagging is a classification task aiming to assign lexical categories to words in a text. Most previous work has concentrated on English and on the usage of supervised methods; however, recently the research focus has shifted to other languages, to unsupervised methods, and to ways to combine taggers. We have developed three part-of-speech taggers for Amharic [16] by training on the corrected version of the manually tagged corpus (of 200,863 words, as described above). All three taggers are based on publicly available systems originally developed for English:

- TnT, Trigrams'n'Tags [10], a fast tagger based on Hidden Markov Models (HMM) which can be trained on different languages / tag sets,
- SVMTool [19], based on Support Vector Machines and high-dimensional vectors, and
- Maximum Entropy Tagging [22], a linear classification of pre-defined weights for which we used the MALLET Java package [21].

All taggers showed clearly worse performance than previously reported results for English and Arabic, but similar to results reported for Hebrew. In particular the taggers had problems with unknown words, which are almost four times as frequent in the 200K words Amharic corpus used as in the (about six times larger) English Wall Street Journal corpus. (A word is taken to be unknown if it can be found in the test data but not the training data.)

When evaluated (using 10-fold cross validation on the same predefined folds) all taggers got comparable results: 92.5–92.8% accuracy on average over the 10 runs on the two reduced tagsets (with 11 and 10 tags each) and 4–7% lower on the full (30 tag) tagset. The SVM-tagger performed slightly better than the others overall, since it had the best performance on unknown words (78.7% and 88.2–88.7% accuracy on unknown words combined with 89.6% and 93.3–93.4% on known words, giving 88.3% accuracy overall on the full tagset and 92.8% on the reduced sets).

TnT gave the best results for known words (90.0% accuracy on the full tagset; 94.0% on the reduced sets), but had the worst performance on unknown words (only 52.1% on the full tagset and 82.1–82.2% on the smaller sets, giving overall performances of 85.6% and 92.6%, respectively). MaxEnt's overall performance was in the middle of the road (87.9% accuracy on the full set and 92.6% on the reduced tagsets).

In a related but independent experiment, Martha and Menzel [23] used TnT and SVMTool to train two taggers with the 30 class tagset on a version of the corpus which basically was the original (non-cleaned) one, but with each part of multiword units tagged rather than the units merged into one as in our experiment. They evaluated (without cross validation) the taggers by training on 95% of the data and testing on the remaining 5%, and also by training on 90 and 80% of the data and testing on the remaining data. Martha and Menzel's best results were 83% accuracy for TnT and 85.5% for SVMTool, in both cases when training on 95% of the corpus (with slightly lower accuracy when training on 90%).

The previously best performing part-of-speech tagger for Amharic was reported on by Sisay [14]. He used Conditional Random Fields and trained on a limited corpus consisting of five news articles (only 1000 words) annotated with the 10 class tagset mentioned above. When evaluated (using 5-fold cross validation) Sisay's tagger obtained an accuracy of 74%, a result obviously caused by the lack of available data at the time.

5. Morphological Analysis

A shallow morphological analyser for Amharic has been developed [7]. The morphological analyser finds all possible segmentations of a given word according to the morphological rules of the language, and then selects the most likely prefix and suffix for the word based on corpus statistics and stem length. The stemmer strips off the prefixes and suffixes, and then tries to look up the remaining candidate stems (or alternatively, some morphologically motivated variants of them) in a dictionary to verify that they are all possible stems of the word. Derivational variants are not handled since they often are found under different dictionary entries.

If entries matching several of the possible stems can be found in the dictionary, the longest matching sequence is selected, while occurrence statistics are used in order to select the most plausible if several possible segmentations of the same length suggested. If, on the other hand, no stem matching the dictionary entries is produced by the system, the candidate stems are modified and the dictionary-lookup is attempted again with the newly created stem candidates. The frequency and distribution of prefixes and suffixes over Amharic words is based on a statistical analysis of the unannotated 3.5 million word Amharic news corpus described above (Section 3.1). The morphological analyser/stemmer had an accuracy of about 85% when evaluated on a limited text consisting of 50 sentences (805 words). In a larger test, performed on Amharic news texts from Walta Information Center (1503 words; of which 1000 were unique), the accuracy was 76.9%.

The morphological analyser has mainly been used for stemming for Information Retrieval, a task also addressed by Nega and Willett [1]. Testing on 1221 random Amharic words and manually assessing the resulting stems, they assessed 95.5% to be “linguistically meaningful”, however, no evaluation of the correctness of the segmentations was given.

The currently most complete morphological processing tool for Amharic is probably Gasser’s HornMorpho 1.0 [18] which uses a finite-state approach to process Amharic nouns as well as Amharic and Tigrinya verbs. No evaluation of the tool’s performance is given in [18]; however, in an invited talk given at the 2009 ACL Workshop on Computational Approaches to Semitic Languages, Gasser indicated that then about 56% of the nouns and verbs got an analysis from the lexical part of the system (i.e., contained known roots), while an additional 12% could be analysed with guesser modules for verbs and verbal nouns.

Similarly, previous work on Amharic morphological processing initially concentrated on verbs (and nouns derived from verbs). Abiyot [8] created a rule-based system with verb root patterns and affixes, while Tesfaye [9] trained probabilistic models to extract stems and affixes with a success rate of 87%, when tested on a small data set (500 words). On a larger scale, Sisay and Haller [15] used a finite-state approach to analyse Amharic verbs, which Saba and Gibbon [3] extended to include all word categories, achieving 88–94% recall and 54–94% precision when tested on biblical text (1620 words). The precision varied widely depending on the word-class, with the lowest figure for verbs (54%). Thus Saba and Demeke [2] describe ways to extend the system to handle 6400 simple verbal stems.

6. Text Classification

In a set of experiments, we have investigated how well a high-level task like text classification can be carried out in Amharic [13, 5, 4]. It has been claimed that stemming (reducing morphological variants to a single stem) is an important preprocessing step that will allow for improved text categorization accuracy, especially for languages with a rich inflectional morphology. In order to investigate this further, we have performed an experiment where the goal has been to see how stemming vs. non-stemming will affect the performance in a 10-class text categorization task. The best accuracy (69.4%) was achieved using the full text as representation. A representation using only the nouns performed almost equally well, confirming the assumption that most of the information required for distinguishing between the various categories actually is contained in the nouns.

However, a classifier using the output of the stemmer of the previous section only performed *almost* on par with the other classifiers. It might seem a bit surprising that the stemmed representation failed to improve the performance of the text classifier on such a morphologically rich language as Amharic. A possible explanation for this lack of improvement is that the main morphological variation in Amharic pertains to the verbs, not the nouns. However, the latter are the main carriers of the information content of the language, and hence the main sources of information for the classifiers, which is also partly confirmed by the results of the experiments.

7. Conclusions

An important idea of the project described in this paper has been to develop the basic computational linguistic resources for Amharic as rapidly as possible and then to use them in order to create new resources as well as to extend and refine the resources themselves. To this end the following steps have been undertaken:

1. An annotated corpus has been both manually and automatically checked and corrected, in several cycles.
2. Part-of-speech taggers, a morphological analyser, and text classifiers have been trained, retrained and refined based on the modified corpus.

The results obtained in the different experiments indicate that this is a feasible approach to creating language processing resources for an under-resourced language such as Amharic. However, the performance of the systems is so far clearly lower than performance on compatible tasks for English. Obviously, this is partially due to the fact that a lot more time and effort has been devoted to the systems for English, but most likely also to the peculiarities of the Amharic language as such.

Tools and resources that can help reduce language barriers and thereby provide people all over the World with improved access to information and services will have beneficial effects for most sectors of society in all countries. Administration, production, trade, distribution, information handling, communication, environment, health care, media, education and research, are just a few of the areas that benefit from the improved accessibility that can be achieved by allowing users to communicate in their own local language. Hence we believe that the tools and resources produced within this project constitute a small, but important contribution to the development of a technology that over time will enable massive social and economic transformations World-wide.

Indeed, the manually annotated corpus (Section 3.3) is already available on the Internet and has been used by other researchers (as exemplified by [23]). The updated and corrected version of the annotated corpus is planned to shortly be released in the same manner. The morphological stemmer is also freely available already, while the part-of-speech taggers and text classifiers still can (and needs to) be improved before being made available for wider use. Here we are currently working on different methods to combine the three statistical taggers and to utilize morphological information in the tagging process.

Even though current performance of some of the language processing tools investigated within the project thus still is lower than what is needed to make them useful in a wider perspective, and in comparison to similar tools for more thoroughly-researched languages, it should be clear that the strategy chosen in this project is a very feasible approach to developing tools and resources for under-resourced languages.

The manually annotated Amharic corpus is in itself quite small (200.000 words), but has already proven its worth as the basis for creating tools and as a vehicle for further research on – and better understanding of – the Amharic language. Aiming to apply machine learning and statistical methods originally developed for other languages in order to rapidly create the first basic language processing tools for an under-resourced language – such as Amharic – has also been shown to be a very useful methodology. However, the tools created are indeed still fairly basic and further extensions are quite clearly needed before truly useful, market-level tools can be released.

8. Acknowledgements

Thanks to everybody who has been involved with or contributed to the project over time, including Atelach Alemu Argaw (Stockholm University); the ELRC (Ethiopian Language Research Center) staff working on the manual corpus annotation, Dr. Girma Demeke, Ato Mesfin Getachew, Ato Samuel Eyassu, and Ato Lemma Nigussie (all previously or presently at Addis Ababa University); Dr. Magnus Sahlgren and Dr. Fredrik Olsson (SICS, Swedish Institute of Computer Science AB); Daniel Yacob (Ge'ez Frontier Foundation); and Ato Negash at the Walta Information Center in Addis Ababa for allowing the news texts to be used for research purposes.

9. References

- [1] Alemayehu, Nega, and Peter Willett, P. 2002. Stemming of Amharic words for information retrieval. *Literary and Linguistic Computing* 17(1):1-17.
- [2] Amsalu, Saba and Girma A. Demeke. 2006. Non-concatinative finite-state morphotactics of Amharic simple verbs. *ELRC Working Papers* 2(3), 304-325.
- [3] Amsalu, Saba and Dafydd Gibbon. 2005. Finite state morphology of Amharic. In *5th Int. Conf. Recent Advances in Natural Language Processing*, pp. 47-51. Borovets, Bulgaria.
- [4] Asker, Lars, Atelach Alemu Argaw, Björn Gambäck, Samuel Eyassu Asfeha and Lemma Nigussie Habte. 2009. Classifying Amharic Web News. *Information Retrieval* 12(3):416-435, New York, New York. Springer Verlag.
- [5] Asker, Lars, Atelach Alemu Argaw, Björn Gambäck and Magnus Sahlgren. 2007. Applying Machine Learning to Amharic Text Classification. In *5th World Congress of African Linguistics*, Cologne, Germany. Rüdiger Köppe Verlag.
- [6] Argaw, Atelach Alemu and Lars Asker. 2005. Web Mining for an Amharic-English Bilingual Corpus. *1st Int. Conf. Web Information Systems and Technologies*, pp. 239–246, Deauville Beach, Florida.
- [7] Argaw, Atelach Alemu and Lars Asker. 2007. An Amharic stemmer: Reducing words to their citation forms. *Computational Approaches to Semitic Languages*, pp. 104–110, Prague, Czech Republic.
- [8] Bayou, Abiyot. 2000. *Design and development of word parser for Amharic language*. MSc Thesis, School of Information Studies for Africa, Addis Ababa University, Addis Ababa, Ethiopia.
- [9] Bayu, Tesfaye. 2002. *Automatic morphological analyser: An experiment using unsupervised and autosegmental approach*. Msc Thesis, School of Information Studies for Africa, Addis Ababa University, Addis Ababa, Ethiopia (2002)
- [10] Brants, Thorsten. 2000. TnT — a statistical part-of-speech tagger. In *6th Conf. Applied Natural Language Processing*, pp. 224–231, Seattle, Washington.
- [11] CIA. 2010. *The World Factbook–Ethiopia*. The Central Intelligence Agency. Washington, DC. www.cia.gov/library/publications/the-world-factbook/geos/et.html (Latest update: 15 January, 2010).
- [12] Demeke, Girma A., and Mesfin Getachew. 2006. Manual Annotation of Amharic News Items with Part-of-Speech Tags and its Challenges. *ELRC Working Papers*, Vol II, Number 1.
- [13] Eyassu, Samuel, and Björn Gambäck. 2005. Classifying Amharic News Text Using Self-Organizing Maps. In *43rd Annual Meeting of the Association for Computational Linguistics: Workshop on Computational Approaches to Semitic Languages*, pp. 71–78, Ann Arbor, Michigan.
- [14] Fissaha, Sisay. 2005. Part of speech tagging for Amharic using conditional random fields. In *43rd Annual Meeting of the Association for Computational Linguistics: Workshop on Computational Approaches to Semitic Languages*, pp. 47–54, Ann Arbor, Michigan.
- [15] Fissaha, Sisay and Joachim Haller. 2003. Amharic verb lexicon in the context of machine translation. In *10th Conf. Traitement Automatique des Langues Naturelles*, vol. 2, pp. 183–192. Batz-sur-Mer, France.
- [16] Gambäck, Björn, Fredrik Olsson, Atelach Alemu Argaw and Lars Asker. 2009. Methods for Amharic Part-of-Speech Tagging. In *12th Conf. European Chapter of the Association for Computational Linguistics: 1st Workshop on Language Technologies for African Languages*, pp. 104–111, Athens, Greece.
- [17] Gambäck, Björn, Fredrik Olsson, Atelach Alemu Argaw and Lars Asker. 2009. An Amharic Corpus for Machine Learning. *The 6th World Congress of African Linguistics*, pp. 89–90, Cologne, Germany. WOCAL.
- [18] Michael Gasser. 2009. *HornMorpho 1.0 User's Guide*. School of Informatics and Computing, Indiana University, Bloomington, Indiana.
- [19] Giménez, Jesús and Lluís Márquez. 2004. SVMTool: A general POS tagger generator based on support vector machines. In *4th Int. Conf. Language Resources and Evaluation*, pp. 168–176, Lisbon, Portugal.
- [20] Hudson, Grover. 1999. Linguistic analysis of the 1994 Ethiopian Census. *Northeast African Studies* 6(3):89-107. Michigan University Press.
- [21] McCallum, Andrew Kachites. 2002. *MALLET: A machine learning for language toolkit*. mallet.cs.umass.edu
- [22] Ratnaparkhi, Adwait. 1996. A maximum entropy model for part-of-speech tagging. In *Empirical Methods in Natural Language Processing*, pp. 133–142, Philadelphia, Pennsylvania.
- [23] Tachbelie, Martha Yifiru and Wolfgang Menzel. 2009. Amharic Part-of-Speech Tagger for Factored Language Modeling. In *7th Int. Conf. Recent Advances in Natural Language Processing*, Borovets, Bulgaria.
- [24] Wintner, Shuly. 2009. Language Resources for Semitic Languages: Challenges and Solutions. In *Language Engineering for Lesser-Studied Languages*, Sergei Nirenburg (ed.), pp. 277-285, IOS Press, Amsterdam, the Netherlands.