Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 3996–4002
Marseille, 11–16 May 2020

Morphological Segmentation for Low Resource Languages

Justin Mott[1], Ann Bies[1], Stephanie Strassel[1], Jordan Kodner[2], Caitlin Richter[2], Hongzhi Xu[2,3], Mitch Marcus[2]
[1]Linguistic Data Consortium, University of Pennsylvania
3600 Market Street, Suite 810, Philadelphia, PA 19104
[2]Department of Computer and Information Science, University of Pennsylvania
3330 Walnut Street, Philadelphia, PA 19104
[3]ICSA, Shanghai International Studies University
1550 Wenxiang Road, Shanghai 201600, China
{jmott, bies, strassel}@ldc.upenn.edu, {jkodner, ricca}@sas.upenn.edu, {xh, mitch}@cis.upenn.edu

Abstract

This paper describes a new morphology resource created by Linguistic Data Consortium and the University of Pennsylvania for the DARPA LORELEI Program. The data consists of approximately 2000 tokens annotated for morphological segmentation in each of 9 low resource languages, along with root information for 7 of the languages. The languages annotated show a broad diversity of typological features. A minimal annotation scheme for segmentation was developed such that it could capture the patterns of a wide range of languages and also be performed reliably by non-linguist annotators. The basic annotation guidelines were designed to be language-independent, but included language-specific morphological paradigms and other specifications. The resulting annotated corpus is designed to support and stimulate the development of unsupervised morphological segmenters and analyzers by providing a gold standard for their evaluation on a more typologically diverse set of languages than has previously been available. By providing root annotation, this corpus is also a step toward supporting research in identifying richer morphological structures than simple morpheme boundaries.

Keywords: low resource languages, morphology, morphological analyzers, morphological segmenters, annotation, data centers

1. Introduction and Motivation

This paper describes a new morphology resource created by Linguistic Data Consortium and the University of Pennsylvania for the DARPA LORELEI Program. The data consists of approximately 2000 tokens annotated for morphological segmentation for each of 9 low resource languages, along with root information for 7 of the languages. The languages annotated show a broad diversity of typological features. This annotated data will support the development of unsupervised morphological segmenters and analyzers by providing a gold standard for their evaluation. In the sections that follow we discuss the goals of the LORELEI program and the morphological segmentation task, related work, the annotation procedures and resulting corpus, research results and future directions. The data described here is included in the LORELEI Representative Language Packs scheduled for publication in the LDC Catalog starting in 2020.

1.1 The LORELEI Program

The DARPA Low Resource Languages for Emerging Incidents (LORELEI) Program aims to advance the capabilities of human language technologies in low-resource languages, with a particular focus on rapidly obtaining situational awareness after the emergence of an unexpected incident such as a natural disaster in a setting where technology does not yet exist for the local language (DARPA, 2014). Linguistic Data Consortium (LDC) has built a variety of linguistic resources for nearly three dozen low resource languages for the LORELEI program. Representative Language Packs – consisting of large volumes of formal and informal monolingual and parallel (with English) text with a variety of manual annotations to support situational awareness, plus a lexicon, grammatical sketch and basic processing tools – are designed to enable research into language universals and cross-language projection and so include some related higher-resource languages. Representative Language Packs have been created for Akan, Amharic, Arabic, Bengali, Farsi, Hindi, Hungarian, Indonesian, Mandarin, Russian, Somali, Spanish, Swahili, Tagalog, Tamil, Thai, Ukrainian, Vietnamese, Wolof, Yoruba, Zulu, plus partial resource packs for Hausa, Turkish, Uzbek and English. Incident Language Packs contain manually labeled evaluation data designed to test system performance on tasks related to situational awareness for one or more surprise languages that remain unknown until the start of each annual evaluation (Strassel and Tracey, 2016). Incident Language Packs have been created for Ilocano, Kinyarwanda, Odia, Oromo, Sinhala, Tigrinya, and Uyghur.

1.2 Overall Research Goals

Morphology analysis is useful as an underlying task for various natural language processing applications. Supervised methods for such analysis require significant training data and suffer from difficulty in transferring tools from one language to another. Unsupervised methods do not suffer from these problems and are therefore better suited for the LORELEI program's research goals of rapidly functioning in a new unanticipated language.

In general, unsupervised morphological analysis has been focused on segmentation rather than feature labeling (Hammarström and Borin, 2011), since completely unsupervised feature discovery given just text input has proven to be very difficult. This analysis task has also mainly been limited to segmentation of text input in orthographies that provide a clear indication of word boundaries of some kind (ibid).

Another important issue for fully unsupervised morphology learning is that systems should be designed to apply to a wide range of linguistic typologies. This means that the models cannot rely on linguistic properties specific to any particular language or language family, but rather must build on language universals and language typological features. In order to observe this important quality for morphology analysis systems, the gold standard data set must cover a wide enough range of language typologies such that language specific approaches will score poorly when evaluated using this data set.

We exclude from our focus here a range of "almost unsupervised" morphology learning tasks that require a small set of seed words or very small training corpus of some kind, such as the work discussed in (Dreyer and Eisner, 2011). Providing the appropriate training sets requires knowledge of the particulars of these various tasks, so avoiding such an approach addresses our research aims more directly.

1.3 Previously Unmet Research Goals

The primary data source for testing unsupervised morphology has been the test sets for the Morpho-Challenge segmentation task (Kurimo et al., 2010), which was held from 2005 to 2010 and whose data remains available for continued research. However, Morpho-Challenge only annotated a few European languages and so it lacks representation of languages with many different morphology typologies, such as infixation and reduplication and other forms of non-concatenative morphology.

Additionally, pure morphological segmentation does not include root identification, i.e., which segment or segments are surface reflections of the underlying root, although this information is often very useful in downstream tasks and from a human perspective seemed natural to include here. For example, the English words *stopped*, *carried* can be segmented as *stopp-ed* and *carri-ed* respectively; root identification would also link the stems *stopp* and *carri* with the root words *stop* and *carry*. Although the segmentation alone does reveal that the two complex words *stopped* and *carried* share the morpheme *-ed*, this is semantically light and thus only provides limited information.

We also wanted to provide a range of languages all segmented according to exactly the same annotation standard, and so needed to take a view of segmentation that goes beyond any language-particular linguistic tradition.

1.4 Annotation to Support Research Goals

The annotated data created in this project allows for the further evaluation of existing segmentation models, e.g., Morfessor (Creutz and Lagus, 2002; Virpioja et al., 2013), but will also support unsupervised learning research on a much wider range of language typologies for which widely-shared test data has not previously existed. The current data set represents a much more typologically diverse set of languages than what is available under Morpho-Challenge. The nine languages in our corpus cover five primary language families (Austronesian: Indonesian, Tagalog; Dravidian: Tamil; Indo-European: Hindi, Russian, Spanish; Niger-Congo: Akan (Twi), Swahili; Uralic: Hungarian), and cover a range of morphological phenomena including suffixation, prefixation, infixation, circumfixation, full and partial reduplication, and vowel harmony.

Further, seven of the nine languages are annotated with root information as well, which can be used to test existing systems that are designed for identifying roots, such as Narasimhan et al. (2015), Luo et al. (2017), and Xu et al. (2018).

Although this data set does not include languages in which templatic morphology plays a sizeable role (primarily of Semitic and Afro-Asiatic language families), it is sufficient to provide a real challenge for the current state-of-the-art in unsupervised morphology.

2. Related Work

2.1 Related Morphological Annotation

The Penn Treebank developed a part-of-speech tagset for English that encoded some morphological distinctions but that did not provide information on morphemes or morphological segments (Marcus et al., 1993).

The Penn Arabic Treebank moved to more complete morphological annotation with the development of a treebank for Modern Standard Arabic (Maamouri et al., 2004). The morphological annotation for Arabic included lemma, full vocalization, transliteration, English gloss, and the identification of constituent morpheme segments manually selected from morphological analyzer output (Kulick et al., 2010). With the parallel development of a treebank and a morphological analyzer for Egyptian Arabic, procedures were developed for additional manual annotation of morphology and morpheme segments for dialectal Egyptian Arabic (Maamouri et al., 2014).

In the three Representative Language Packs produced by LDC under the pre-LORELEI BOLT program (DARPA 2019), namely Turkish, Hausa and Uzbek, the approach to morphological annotation was tightly integrated with creation of language-specific analyzers at LDC (Kulick and Bies, 2016). One additional type of morphological annotation appears in the Turkish and Uzbek Representative Language Packs: morpheme alignment. This task was designed to identify translational correspondence at the morpheme level in parallel text.

2.2 Research in Unsupervised Morphology Learning

The Morpho-Challenge tasks and their associated data sets contributed to important work on unsupervised morphology learning, including the Morfessor family of models. The Morfessor baseline system (Creutz and Lagus,

2002; Virpioja et al., 2013) served as the baseline in these tasks. Further extended versions of Morfessor also exist, such as Morfessor CatMAP (Creutz and Lagus, 2007), Morfessor FlatCat (Grönroos et al., 2014), etc. As discussed above, most of these systems are only designed for identifying morpheme boundaries. The annotation discussed here is intended primarily to move this stream of work forward.

Some more recent systems focus on identifying morphologically related word pairs, such as (*stop*, *stopped*), and then transform the output to morpheme segmentations so they are compatible with the available segmentation based evaluation. Such work includes Schone and Jurafsky (2001); Narasimhan et al. (2015); Soricut and Och (2015); Luo et al. (2017) and Xu et al. (2018). These systems have the advantage of finding morphologically related word pairs, which is equivalent to finding roots for complex words.

Another stream of completely unsupervised work primarily aims at discovering morphological paradigms (Parkes et al., 1998; Goldsmith, 2001; Chan, 2006). Xu et al. (2018) discover such paradigms, but primarily use them for improving a probabilistic segmentation model. The annotation we report on here does not attempt to determine paradigm information for a wide set of word roots, which is a somewhat different task. While providing a good set of test data for such work is important, only a few systems to date have attempted this task.

A limitation of this research paradigm, of course, is that systems that are directly designed to identify morpheme boundaries do not provide more information than the morpheme itself. Thus, it is not possible to tell the exact morphological structure of complex words, including how these complex words are derived from simpler words and what kind of morphological features the morphemes are corresponding to such as prefixes, suffixes, etc.

One important shortcoming of the existing research discussed above which was evaluated using the Morpho-Challenge test sets is that the evaluation provided relatively small motivation to deal with types of morphology other than simple prefixation and suffixation.

3. Annotation Scheme

3.1 Annotation Requirements

The requirements for morphological segmentation annotation were informed both by the research goals discussed above and by practical considerations of budget and timeline. To address these requirements, the annotated data:

- Should cover a wide range of morphological typologies.
- Should cover a wide range of language families.
- Should be in an orthography that in general indicates word boundaries.
- Should be in an orthography that allows easy division of words in segments.
- Should allow a test set for each language sufficient to measure and thus support progress in development of unsupervised morphology learning for the medium term.
- Should be affordable within the limits of an ongoing research program (LORELEI) whose research goals are significantly broader than those described here.
- Should be annotated quickly and with reasonable accuracy.
- Should mark surface segments that realize the semantic root.

These requirements constrain what languages could be annotated and what annotation could be performed. For reasons of cost and annotator availability, the languages chosen also needed to be a subset of those which were already being addressed within the LORELEI program. We limited ourselves to languages whose orthography was in an alphabet or abugida, rather than a syllabary or logosyllabic (like Chinese). We also limited ourselves to languages where word spaces are used.

The annotation performed needed to be quite light weight to allow the cost per language to be low, and thus to maximize the number of languages annotated. This led to an annotation scheme which focused on segmentation as the primary mechanism, since it is an annotation that can be performed entirely on text with very limited training or input from a linguist. Because the training requirements are light, it can be rapidly deployed to new data sets and new languages as the need arises.

Because the annotation scheme serves the purpose of evaluating unsupervised morphological segmenters and analyzers, and the accuracy of current systems is below 70% in F1 measure for most languages, a relatively small amount of data will be useful to support measurement of meaningful steps in research progress in the short term. We determined that 2000 tokens per language would provide a reasonable minimum for this purpose while still allowing room in the budget and timeline for a range of language types to be annotated.

Additional data could be added per language in the future to make progress toward unsupervised morphology discovery; this may become necessary as accuracy levels reach the point that particular small phenomena need to be represented in the data set. For the moment, we have privileged more languages over more data to attempt to cover diverse typologies.

To test that the planned morphological segmentation task would enable rapid, accurate annotation, the annotation scheme was originally developed by the University of Pennsylvania team and applied to English, Spanish, Faroese, and Korean before expanding to a broader set of LORELEI languages in collaboration with LDC. The first three languages were annotated with multiple annotations performed to test accuracy, and Korean was then annotated by a linguistics PhD student at UPenn. The annotation scheme evolved throughout this work, thus we do not report accuracy figures here.

3.2 Developing Annotation Guidelines

Starting with the task as initially defined by the UPenn team, LDC developed procedural annotation guidelines for this task that apply the principles of morphological segmentation annotation discussed below. Language-independent principled specifications suitable for use by annotators were developed by LDC, and appropriate language-specific morphological paradigms and other specifications for each target language were included in an appendix for each language.

Segmentation is a minimal annotation scheme that can capture the patterns of a wide range of languages and can be performed reliably by annotators, and which can easily cover suffixation, prefixation, infixation, circumfixation, and reduplication. This scheme can be extended to include simple mutations. We used angle brackets to mark forms which annotators viewed as mutated from a base, e.g., *br<a>ng* (from *bring*).

We do not mark suppletion, so *brought* is annotated as a single morpheme with no further detail. Derivational and inflectional morphology were not distinguished in the segmentation. The annotation scheme was not designed with compounding in mind, but segmentation of compounds fell out naturally in the annotation.

Finally, we also added a marking of surface root forms where time permitted, on a per language basis.

4. Corpus Description

4.1 Languages and Data Volumes

The languages for morphological segmentation annotation were selected based on criteria targeting the annotation of a variety of linguistic features and language types across the 9 annotated languages. Table 1 shows the features and data volume for each language.

Language	ID	Transliteration	Root	Tokens
Akan (Twi)	aka	N	N	2048
Hindi	hin	Y	Y	2028
Hungarian	hun	N	Y	2027
Indonesian	ind	N	Y	2035
Russian	rus	Y	Y	2050
Spanish	spa	N	Y	2050
Swahili	swa	N	Y	2023
Tagalog	tgl	N	Y	2001
Tamil	tam	Y	N	2028

Table 1: Annotated Languages and Data Volume

4.2 Morphological Features Covered

The languages for this data set were downselected from the full set of LORELEI Representative Languages to cover a diversity of typological features. Table 2 illustrates the targeted morphological features for the annotated languages, and Table 3 provides an annotated example for each language.

Feature	Language(s)
Robust case system	Hungarian, Russian, Tamil
Infixation	Indonesian, Tagalog
Noun class prefixes	Akan (Twi), Swahili
Agglutination	Swahili, Tamil
Circumfixation	Indonesian
Reduplication	Akan (Twi), Indonesian, Tagalog

Table 2: Morphological Features of Annotated Languages

lang	token	transliteration	Annotation
aka	apolisifoɔ		a polisi foɔ
hin	किए	kie	k<i> e
hun	legerősebb		leg erő sebb
ind	memberikan		mem beri kan
rus	медицинских	meditsinskih	meditsin sk ih
spa	puedan		p<ue>d an
swa	ilipozungumza		i li po zungum z a
tgl	kumakain		k um a kain
tam	மரங்கள்	marangal	mara<n> gal

Table 3: Examples of Annotated Data

4.3 Data Selection

LDC developed token selection criteria and procedures in consultation with the University of Pennsylvania team that allowed the annotation of representative tokens from each targeted language, and performed human annotation of each selected token. Approximately 3000 tokens per language were selected from the Representative Language Pack's 25Kw situational awareness data set for that language; this data had been previously annotated for multiple tasks within LORELEI. The top ten most frequent tokens were excluded from the annotation pool, as were infrequent tokens and non-linguistic tokens (e.g., URLs). Out of this pool of candidate tokens for annotation, approximately 2000 tokens were manually annotated in each targeted language. Tokens not suitable for annotation were manually excluded.

Tokens from non-Latin script languages were transliterated using the transliterator provided in the Representative Language Pack for that language, and the manual annotation was performed on the transliterated tokens. In part, this was done because otherwise the segmentation annotation would have been cumbersome in the languages written in abugidas (i.e., Hindi and Tamil).

For each language, there is also a concordance file. The file consists of five tab delimited columns: doc_id, token_id, token, start_offset, end_offset. The doc_id information matches the file name in the Representative Language Pack for that language. For each annotated token, the concordance consists of every occurrence of that token string appearing in the situational awareness data set for that language.

Because the annotation of the tokens is done in isolation (not in the context of full document text), the concordance will include tokens that may appear in a different context from the token that was annotated. It is therefore the case that not every instance of the token in the concordance

would necessarily have the same morphological analysis as the annotated token.

4.4 Annotation Process

The annotation was done in two stages. A first pass annotation was done in 2018 and a subsequent quality control (QC) pass performed in 2019. In each stage, the annotation was performed by a non-linguist native speaker annotator paired with an LDC linguist managing the overall annotation effort.

For the first pass, annotators were presented with the annotation guidelines along with the language-specific appendix for their language. Annotators then completed a brief practice set that was reviewed by the LDC linguist. Segmentation annotation was performed on the selected word list, with annotators working in tandem with the LDC linguist. Annotators were allowed to reject tokens for being out of language or otherwise being unsuitable for annotation.

For the quality control pass, a manual review to flag errors was performed on the first pass annotation. Similar strings with different segmentation were flagged for manual review. A histogram of the most commonly annotated segments for each language was created, and those segments in the annotation were examined to identify variations that were flagged for manual annotator review.

Annotators who worked on quality control corrections were trained on the goals of the correction and QC task, with an emphasis on consistency. QC annotators were also lightly trained on the annotation guidelines for the morphological segmentation task (including the language-specific appendix for their language). Finally, root identification was performed on languages in this pass, time permitting.

5. Research Progress

5.1 Progress as Supported by Annotation

With the wide coverage of language families by our data set, we can explore how existing segmentation models perform on a wide range of languages. We compare four existing models: Morfessor (Virpioja et al., 2013), Morpho-Chain (Narasimhan et al., 2015), ILP (Luo et al., 2017), ParaMA (Xu et al., 2018) systems. Morpho-Chain and ILP were run both with word vectors (+vec) and without. The word vectors are of 200 dimensions and are trained with the word2vec model based on the LORELEI language pack. The results are shown in Table 4.

Lang	Morf	MC	ILP	MC+vec	ILP+vec	ParaMA
Aka	0.6337	0.6504	0.6461	0.4986	0.5308	0.5306
Hin	0.2588	0.3594	0.3469	0.4945	0.5054	0.5866
Hun	0.4078	0.5319	0.5339	0.6228	0.6196	0.5327
Ind	0.5322	0.4998	0.4975	0.5614	0.6219	0.4692
Rus	0.3478	0.4924	0.4937	0.4270	0.4504	0.4586
Spa	0.2507	0.4985	0.5027	0.0514	0.0346	0.4057
Swa	0.4320	0.4306	0.4089	0.2020	0.1898	0.3435
Tgl	0.5254	0.4846	0.4708	0.4304	0.4391	0.4119
Tam	0.2376	0.2937	0.2916	0.3414	0.3359	0.3966
Avg	0.4028	0.4712	0.4658	0.4033	0.4142	0.4594

Table 4: Experimental results in F1 measure on testing existing morphological segmentation systems.

From these results, it is clear that the limited amount of test data is not a significant barrier to measuring research progress. Results for these systems are generally in the range of 0.2 to 0.6 F1, with the maximum over all systems over all languages under 0.65. It is reasonable to guess that the data quantities provided will suffice for the foreseeable future. It is also clear that different systems perform better on different languages, with no one system showing top performance across all languages. Similarly, different systems have significant difficulty relative to average performance on different languages. This may well indicate an (implicit) bias within each system towards different language typologies.

6. Conclusions

6.1 Mutual Benefits of Annotation and Research Progress

This dataset will stimulate the field of unsupervised morphology learning by encouraging research on a much wider range of morphological typology and a much richer set of language families than can presently be evaluated using the dominant current test set from the MorphoChallenge competitions. By including root annotation for seven of the nine languages, this data also provides a step toward supporting research in identifying richer morphological structures than simple morpheme boundaries.

6.2 Future Work

We are well aware of the limits of the current corpus, but we anticipate that this data will stimulate the research that will ultimately make this corpus obsolete.

As new unsupervised methodologies allow big error reductions over current systems, larger test sets will certainly be necessary.

Developing annotation standards for templatic morphology that can be applied by untrained annotators across the wide range of Semitic and Afro-Asiatic languages which utilize templatic signalling of morphology is a significant remaining challenge, beyond the relatively simple cases of Modern Standard Arabic and modern Hebrew. A further question for such languages is whether evaluation should

be done on vocalized texts (e.g., Arabic texts with *harakat* (vowel marks)), not used by literate readers in some Semitic languages in particular, or in the unvocalized version.

In the future, data sets which provide much deeper morphological analyses will be needed to spur research. The field will need quite different schemes for evaluating the system accuracy, and some kind of consensus across researchers as to what such an annotation should include is needed. Providing morphological features will ultimately be useful, we would hope, but again, a consensus would need to emerge as to what a completely unsupervised system should be able to discover.

The work we report here is a first step in this direction.

These resources discussed in this paper been released in language packs to participants in the LORELEI program, and will be included in the LORELEI Representative Language Packs scheduled for publication in the LDC Catalog starting in 2020.

7. Acknowledgments

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-15-C-0123 and Contract No. HR0011-15-2-0023. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of DARPA.

8. Bibliographical References

Mathias Creutz and Krista Lagus. 2002. Unsupervised discovery of morphemes. In Proceedings of the ACL02 workshop on Morphological and phonological learning. Volume 6, pages 21–30. Association for Computational Linguistics.

Mathias Creutz and Krista Lagus. 2007. Unsupervised models for morpheme segmentation and morphol-ogy learning.ACM Transactions on Speech and Language Processing (TSLP), 4(3):1–34.

DARPA. 2014. Broad Agency Announcement: I2O Low Resource Languages for Emergent Incidents (LORELEI). Defense Advanced Research Projects Agency, DARPA-BAA-15-04.

DARPA. 2019. Broad Operational Language Translation (BOLT) (Archived). Retrieved from https://www.darpa.mil/program/broad-operational-language-translation .

Markus Dreyer and Jason Eisner. 2011. Discovering morphological paradigms from plain text using a Dirichlet process mixture model. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 616–627.

Jennifer Garland, Stephanie Strassel, Safa Ismael, Zhiyi Song, Haejoong Lee. 2012. Linguistic Resources for Genre-Independent Language Technologies: UserGenerated Content in BOLT. Proceedings of LREC 2012: 8th International Conference on Language Resources and Evaluation, Istanbul, May 21-27.

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

Stig-Arne Grönroos, Sami Virpioja, Peter Smit, Mikko Kurimo. 2014. Morfessor FlatCat: An HMM-based method for unsupervised and semi-supervised learning of morphology. In Proceedings of the 25th International Conference on Computational Linguistics. Pages 1177–1185, Dublin, Ireland.

Harald Hammarström, Lars Borin. 2011. Unsupervised Learning of Morphology. In Computational Linguistics, Volume 37, Number 2, June 2011,

Seth Kulick, Ann Bies. 2016. Rapid Development of Morphological Analyzers for Typologically Diverse Languages. Proceedings of LREC 2016: 10th International Conference on Language Resources and Evaluation, Portorož, May 23-28.

Seth Kulick, Ann Bies, Mohamed Maamouri. 2010. Consistent and Flexible Integration of Morphological Annotation in the Arabic Treebank. In Proceedings of LREC 2010: 7th International Conference on Language Resources and Evaluation, Valletta, Malta, May 17-23.

Mikko Kurimo, Sami Virpioja, Ville Turunen, Krista Lagus. 2010. Morpho Challenge competition 2005-2010: Evaluations and results. In Proceedings of the 11th Meeting of the ACL-SIGMORPHON, ACL 2010, pages 87–95, Uppsala, Sweden, 15 July 2010.

Constantine Lignos. 2010. Learning from unseen data. In Proceedings of the Morpho Challenge 2010 Workshop, pages 35–38.

Mohamed Maamouri, Ann Bies, Tim Buckwalter, Wigdan Mekki. 2004. The Penn Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus. In Proceedings of NEMLAR International Conference on Arabic Language Resources and Tools, Cairo, Egypt, September 22-23.

Mohamed Maamouri, Ann Bies, Seth Kulick, Michael Ciul, Nizar Habash and Ramy Eskander. 2014. Developing an Egyptian Arabic Treebank: Impact of Dialectal Morphology on Annotation and Tool Development. In Proceedings of LREC 2014: 9th Edition of the Language Resources and Evaluation Conference, Reykjavik, Iceland, May 26-31.

Mitchell P. Marcus, Beatrice Santorini and Mary Ann Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. In Computational Linguistics, Volume 19, Number 2, June 1993, Special Issue on Using Large Corpora: II.

Karthik Narasimhan, Regina Barzilay, and Tommi Jaakkola. 2015. An unsupervised method for uncovering morphological chains. Transactions of the Association for Computational Linguistics, 3:157–167.

Cornelia Parkes, Alexander M. Malek, and Mitchell P.Marcus. 1998. Towards unsupervised extraction of verb paradigms from large corpora. In Proceedings of the Sixth Workshop on Very Large Corpora (COLING-ACL).

Patrick Schone and Daniel Jurafsky. 2001. Knowledge-free induction of inflectional morphologies. In Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies, pages 1–9.

Radu Soricut and Franz Och. 2015. Unsupervised morphology induction using word embeddings. In

Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1627–1637.

Strassel, S., and Tracey, J. (2016). LORELEI Language Packs: Data, Tools, and Resources for Technology Development in Low Resource Languages. In Nicoletta Calzolari (Conference Chair), et al., editors, Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pages 3273-3280, Portoroz, Slovenia, May 23-28. European Language Resource Association (ELRA).

Sami Virpioja, Peter Smit, Stig Arne Grönroos, Mikko Kurimo et al. 2013. Morfessor 2.0: Python implementation and extensions for morfessor baseline.

Hongzhi Xu, Mitchell Marcus, Charles Yang, and Lyle Ungar. 2018. Unsupervised morphology learning with statistical paradigms. In Proceedings of the 27th International Conference on Computational Linguistics, pages 44–54.

9. Language Resource References

Linguistic Data Consortium. (2015a). BOLT LRL Hausa representative language pack v1.2. LDC2015E70.

Linguistic Data Consortium. (2015b). BOLT LRL Turkish representative language pack v2.2. LDC22014E115.

Linguistic Data Consortium. (2016). BOLT LRL Uzbek representative language pack v1.0. LDC2016E29.