Corpus Building for Low Resource Languages in the DARPA LORELEI Program

Jennifer Tracey, Stephanie Strassel, Ann Bies, Zhiyi Song, Michael Arrigo, Kira Griffitt, Dana Delgado, Dave Graff, Seth Kulick, Justin Mott and Neil Kuster
Linguistic Data Consortium
3600 Market Street, Suite 810
Philadelphia, PA 19104
{garjen|strassel|bies|zhiyi|marrigo|kiragrif|
danafore|graff|kulick|jmott|neilkus@ldc.upenn.edu}

Abstract

We describe corpora for the LORELEI (Low Resource Languages for Emergent Incidents) Program, whose goal is to build human language technologies to provide situational awareness during emergent incidents, with a particular focus on low resource languages. Incident Language packs are used for system development and testing in machine translation, entity disambiguation and linking, and the "situation frame" task, which requires aggregation of information about the emergent incident. Incident languages, as well as the incidents themselves, remain unknown until the evaluation begins, and no labeled training data is provided; systems developers must rapidly adapt technology for the incident language and return initial results within 24 hours. Given this surprise language evaluation scenario, Representative Language packs are designed to support research into cross-language projection and language universals rather than to provide training data. They contain large volumes of monolingual and parallel text, basic annotations, lexical resources and simple NLP tools for 23 languages selected for typological diversity and coverage. We discuss the creation of the LORELEI language packs with a special focus on resources for machine translation, as well as techniques for maintaining consistency across the language packs.

1 Introduction

The DARPA (Defense Advanced Research Projects Agency) LORELEI Program aims to improve the performance of human language technologies capable of providing situational awareness in the context of a specific natural disaster or other emergent incident, with a particular focus on low resource languages for which existing natural language processing (NLP) technology including machine translation is insufficient to support the use case. Systems are required to process information about topics, entities, relations and sentiment, where both the incident language(s) and the incident type(s) remain unknown until the evaluation starts, and where initial system output is due in just 24 hours.

Specialized data is crucial to achieving these ambitious goals. Linguistic Data Consortium (LDC) has created two types of linguistic resources to support training, development and evaluation of machine translation (MT) and other language technologies for LORELEI: Incident Language (IL) packs and Representative Language (RL) packs. Incident Language packs are designed for LORELEI system development and testing. Systems are evaluated against a human gold standard reference for three tasks: machine translation, entity disambiguation and linking, and "situation frame", which requires aggregation of basic information about the needs and issues resulting from the emergent incident. MT output is also subject to human assessment to gauge its utility for of the situation frame task.

Along with a blind test set, Incident Language packs include a small "rapid adaptation" training set containing the type of found data that might be discoverable for a low resource language at the outset of an incident. We have created Incident Language packs in seven languages to date, with two more currently in progress to support the final LORELEI program evaluation in summer 2019. The IL languages appear in Table 1.

Kinyarwanda	Uyghur
Oromo	Uzbek
Sinhala	IL11 (undisclosed)
Tigrinya	IL12 (undisclosed)
Ukrainian	

Table 1: LORELEI Incident Languages

Representative Language packs contain resources in 20 languages that have been selected to provide broad typological coverage, with languages ranging from higher resourced (Spanish) to very low resourced (Akan). Partial language packs exist for three additional languages. Because evaluation languages remain unknown to system developers until the start of the test period, RL packs are designed to support research into utilization of language universals and projection from related-language resources, rather than serving as training data tailored to particular evaluation tasks in a pre-specified language. Each RL pack contains large volumes of monolingual and parallel text, along with smaller amounts of manual entity annotation and linking, light semantic role labeling, document-level labeling of situational needs and issues, as well as a lexicon and basic tools such as tokenizers and sentence segmenters, plus a grammatical sketch for the language. Some RL packs include supplemental morphological or syntactic resources. Every RL pack also shares a common set of English documents translated into the RL; when this set is combined across all RLs it comprises a 21-way parallel corpus. The RL languages appear in Table 2.

Akan (Twi)	Swahili
Amharic	Tagalog
Arabic	Tamil
Bengali	Thai
Farsi	Vietnamese
Hindi	Wolof
Hungarian	Yoruba
Indonesian	Zulu
Mandarin	English (partial)
Russian	Hausa (partial)
Somali	Turkish (partial)
Spanish	

Table 2: LORELEI Representative Languages

In the sections that follow we discuss the process used to create the LORELEI language packs, with a particular focus on resources to support machine translation research. We also discuss strategies for maintaining compatibility and consistency in data collection, translation and annotation efforts across all LORELEI languages.

2 Monolingual Test, Parallel Text and Lexicons

LORELEI RL and IL language packs contain a number of components specifically designed to support machine translation research, including large volumes of monolingual and parallel or comparable text as well as rich lexical resources.

2.1 Monolingual Text

Both Representative and Incident Language Packs contain large volumes of monolingual text, primarily focusing on data in the LORELEI domain of emergent situations like natural disasters, and spanning a range of genres from formal news to informal social media, blogs and discussion forums to reference materials like Wikipedia. The minimum target for monolingual text in the RLs was 2 million words; actual data yields ranged from over 1.25 billion words on the high end (Russian) to 600,000 words on the low end (Wolof, the only language to fall below the minimum target). IL minimum targets were lower, and final data volumes ranged from 3 million words (Oromo) to 27 million words (Uyghur). Reaching the minimum data volume targets for ILs proved to be a particular challenge, especially for some genres; this was exacerbated by the need for the IL test sets to be primarily comprised of documents about the particular test incident(s). We relied heavily on IL native speakers to use creative search techniques to find test incident data, and often needed to stretch the boundaries of traditional genre definitions to satisfy minimum IL data volume targets.

The data collection process involved seeding the corpus with documents known to be in the LORELEI domain generally (for RLs) or about the particular test incident(s) (for ILs). Native speakers for each language searched the web for suitable sources in their language, selecting particular documents with incident- or domain-relevant topics as well as full websites that contain large volumes of appropriate general content for that language. Incident keywords were also used to identify additional in-domain documents for each language. Each website or document selected for inclusion in the corpus was then harvested using an extension of LDC's web collection infrastructure first developed in the DARPA BOLT Program (Garland et. al. 2012). Harvested text was tokenized and sentence-segmented using LORELEI tools designed for cross-language consistency, supplemented with existing open source tools where necessary, and encoding was standardized to UTF-8. The Google CLD2 language detector was used to filter out harvested text not in the target language. CLD2 performance varied considerably by language and genre; moreover, data for many languages contained some degree of code switching and orthographic variation. Therefore, automatic language ID was intended to locate pervasive problems with a data source, rather than detect every case of non-target text in the corpus. Given this, documents subject to manual translation and annotation received an additional manual audit pass to verify language, content and domain relevance.

All collected sources were also reviewed for Intellectual Property Rights issues prior to distribution, and where necessary language packs include pointers to the original data rather than downloaded/processed data. Language packs include utilities for corpus users to download and process such data, to ensure that they end up with the same versions of the data LDC used throughout our data pipeline.

2.2 Parallel and Comparable Text

Representative Language packs contain a minimum of 1 million words of parallel text: 900,000 words of RL source data translated into English, and a common set of 100,000 words translated from English into every RL. The 900,000 word set was drawn from the monolingual text collection for each language, and was designed to contain roughly equal proportions of data from formal news sources and from informal genres like blogs and social media, though the actual distribution varied by language. The common set of 100,000 English words translated into every RL contained four components: approximately 50% general English news documents, 25% LORELEI-domain English news documents, with the remaining 25% consisting of a phrasebook and elicitation corpus originally developed for the REFLEX (Research on English and Foreign Language Exploitation) Program and subsequently updated for LORELEI (Alvarez et al., 2006). Because the same 100,000-word English set was translated into all 20 RLs, the result is a 21-way parallel corpus spanning a broad range of language families and typologies.

We used three methods to produce parallel text for the Representative Languages: 1) scraping parallel text from the web; 2) obtaining translations through crowdsourcing; and 3) obtaining translations from translation vendors. This combination of methods resulted in translations of varying quality and quantity across languages, but the goal was always the same: to produce sentence-aligned content-accurate translations.

Wherever possible translation targets were met by scraping existing parallel text from the web. In addition to harvesting parallel text sources identified by native speakers, we used BITS (Ma and Liberman, 1999) to locate additional sources of parallel text from the web. BITS scans a list of potential parallel websites, downloads content from those websites and uses translation lexicons constructed for LORELEI to perform language ID on the individual webpages and identify any that are translations of one another. The document pairs are then sentence aligned using Champollion (Ma, 2006), which calculates similarity scores between tokenized segments from both languages to reach the optimal alignment.

When found parallel text was insufficient to meet data volume targets, we turned to crowdsourcing, using two platforms: Amazon Mechanical Turk (https://www.mturk.com/) and CrowdTrans (https://crowdtrans.com/), a platform first developed under LORELEI. Initial crowdsourcing efforts focused on translation of English news sources into RLs, with good yields for Spanish, Russian and Arabic. Subsequent efforts were limited to languages with very large pools of crowd workers, namely Hindi and Benglai, and focused on RL-into-English data. Translation proceeded one segment at a time in order to maintain accurate sentence alignment across languages. Segments resulting in at least 3 crowd translations were also subject to a crowdsourced best-to-worst ranking task for additional quality control. Within the CrowdTrans platform we also used native speaker Community Managers to vet workers before translation to improve the overall quality.

When the combined yield from crowdsourcing and found parallel text did not satisfy the target data volume for a given language, we relied on experienced LDC translation vendors who translated whole documents, maintaining sentence-to-sentence correspondence across the language pairs. Unsurprisingly, we relied most heavily on translation vendors for the lowest-resourced languages, where there was very little existing parallel text on the web and where there were too few workers to make crowdsourcing viable.

Taken together, the LORELEI Representative Language packs provide nearly 42 million words of parallel text, of which 68% came from found data, 5% from crowdsourcing and 27% from translation vendors. Figure 1 shows the relative use of each method across the Representative Languages. Note that for some languages like Chinese and Arabic, existing high quality vendor translations were already available from prior DARPA language programs, so very little new translation was produced using any method.

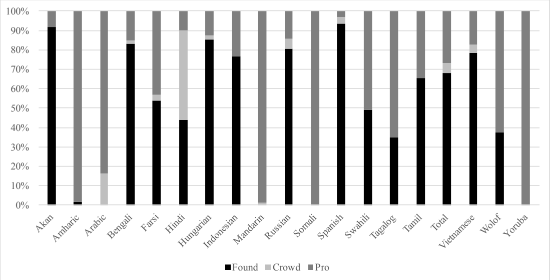

Figure 1: Translation Methods for LORELEI Representative Languages

Incident Language packs include up to 75,000 words of IL data with high quality (one-way, two-way or four-way) manual English translations for system evaluation. They were designed to include an additional 300,000 words of found parallel text for rapid system adaptation into the IL; ideally at least some of this adaptation data is relevant to the evaluation incident or at least relevant to the LORELEI domain. When compared with the Representative Languages, LORELEI ILs are very low resource, and in many cases sufficient volumes of parallel text simply do not exist on the web to satisfy the 300,000 word target. In these cases we provided much larger volumes of comparable text instead. To produce IL comparable text we harvested large volumes of data in both the IL and English, including multilingual data from the same website where possible. We created multilingual clusters following Kutuzov et al. (2016), with document vectors obtained using the method described in Arora et al (2017). The resulting clusters were then augmented by assigning individual Incident Language documents to English-only clusters whose centroid was maximally similar to the IL document. Native speakers reviewed the results to prune, merge and split clusters as needed to improve the overall quality.

2.3 Lexicons

Each Representative Language pack includes a lexicon encompassing an inventory of at least 10,000 headwords/lemmas with part-of-speech, English gloss, and optionally (where appropriate and available) morphological information. The lexicon is comprised of found resources like existing online dictionaries, etc., with some manual effort by native speakers to create new entries as needed to ensure adequate coverage, focusing effort on high frequency tokens missing from the found resources.

For several languages, extensive morphological information is included in a separate wordforms table indexed to the entries in the lexicon. For Arabic, morphological information comes from the Penn Arabic TreeBank (Kulick, et al., 2010); for Amharic, Farsi, Hungarian, Russian, Somali, Spanish and Yoruba, morphological information comes from the Unimorph Project (Kirov, et al., 2018).

For Incident Languages, a custom lexicon is not created, but pointers to available online monolingual and bilingual dictionaries are provided, and where terms of use permit redistribution, the dictionaries are included in the corpus. ILs also include other found grammatical resources like gazetteers, grammars and primers.

3 Annotation Resources

In addition to monolingual and parallel text, LORELEI Language Packs contain several types of manual annotation.

3.1 Entity Annotation

LORELEI Language Packs include three kinds of entity annotation resources: Simple Named Entity, Full Entity, and Entity Linking. In Simple Named Entity (SNE) annotation, text is labeled for names of persons, organizations, locations/ facilities, and geopolitical entities. Full Entity (FE) annotation adds nominal and pronominal mentions of the same types, as well as titles for person entities (such as job titles); it also adds document-level entity coreference. In Entity Linking, named entity mentions are linked to a reference knowledge base developed for LORELEI based on existing external resources. A total of 75,000 words per representative language was labeled for SNE, while an additional 25,000 words was labeled for FE and Entity Linking. Incident Language Pack test sets were also labeled for SNE and EDL.

3.2 Semantic Annotation

Representative Language Packs contain two semantic annotation types developed to support LORELEI research and evaluation: Simple Semantic Annotation (SSA) and Situation Frame (SF). Both SSA and SF label basic information relevant to humanitarian aid and disaster relief scenarios. Situation Frame annotation directly corresponds to the LORELEI SF evaluation task, with a focus on the kind of information that monolingual English-speaking mission planners might require in order to direct a response to an incident as it unfolds. For each document annotators identify the kinds of needs that exist in each location, as well as issues such as civil unrest that might affect the provision of humanitarian assistance, along with the entities involved in the incident. Annotations are at the document level and do not involve extracting specific text extents to justify each Situation Frame. Each annotated frame also includes information about the status, scope, and severity of the needs and issues, as well as sentiment or emotion expressed toward them. SSA represents a more general approach to semantic annotation, labeling basic information about physical events and disaster-relevant situations, their participants, and their locations, with annotations tied to specific text extents in the data. SF appears in both RL and IL packs, while SSA appears only in RL packs.

3.3 Morphosyntactic Annotation

Two types of morphological and syntactic annotation appear in the Representative Language Packs. A 10,000-word subset of the data labeled for both Full Entity and Simple Semantic Annotation is further annotated to identify maximal, non-overlapping Noun Phrases (NPs). Annotators follow surface syntactic structure, applying constituency tests to determine where to mark NPs. After the first 10 RLs were created, a programmatic decision was made to put more annotation effort toward entity and Situation Frame annotation, and so NP annotation was not added to the remaining RLs.

Morphological segmentation was also performed for nine languages; these languages were selected to include a variety of morphological features including robust case marking systems and noun class systems, the use of infixes, circumfixes, reduplication, etc. The nine languages selected were Akan, Hindi, Hungarian, Indonesian, Russian, Spanish, Swahili, Tagalog, and Tamil. For each of these languages, 2000 tokens were segmented at morpheme boundaries, along with markup to indicate substitution (as in *give/gave* in English).

3.4 Parallel Annotation Set

As noted above, all Representative Language packs share a common set of documents translated from English into the RL. From this common set, a smaller 2000-word set was selected for parallel annotation in both the original English and the RL translation. This means that the same translated-from-English document set has been annotated in English and in each of 20 RLs, for all of the following tasks: Simple Named Entity, Full Entity, Entity Linking, Simple Semantic Annotation, and Situation Frame. This data has also been labeled for Noun Phrase Chunking in the 10 RL languages where that task was completed.

4 Grammatical Sketches and Tools

Beyond monolingual text, parallel text and annotations, Representative Language Packs also include grammatical sketches focusing on paradigms and basic grammatical descriptions intended to convey practical information about how to work with the language, rather than deep theoretical discussions or nuanced explication of exceptional cases. Sketches for all languages follow a single template, and include basic information about the language (classification, ISO code, word order, etc.), orthography (characters, variation, word boundaries, etc.), encoding (Unicode chart, etc.), morphology, syntax, and specialized subgrammars for personal names, locations, and numbers, as well as information about variation and references to in-depth grammars. IL packs do not include a customs grammatical sketch, but they do include pointers to grammatical resources about the IL, in the Incident Language and/or in English.

LORELEI RL packs also include basic NLP tools, intended to provide baseline-level performance rather than state-of-the-art. These tools include a transliterator for languages written in non-Roman scripts, tokenizers, sentence segmenters, and named entity taggers. For languages with whitespace-delimited words, we create a custom tokenizer that operates on a series of regular expressions that dictate how to tokenize while preserving web-text artifacts such as hashtags and URLs as single tokens. For languages that do not use whitespace at word boundaries, we rely on existing widely-used tokenizers. Sentence segmentation is performed using an implementation of the Punkt algorithm based on the version found in NLTK (Kiss et. al. 2006). The named entity tagger is a custom conditional random field-based named entity tagger for each RL, trained on the named entity annotations described above.

5 Evaluation Resources for Machine Translation

The primary machine translation evaluation for LORELEI relies on one-way, two-way or four-way gold standard manual translation of an incident-focused test set for each IL. This test data is supplemented by two additional MT evaluation resources: Assessment and HyTER.

5.1 Assessment of MT Output for the Situation Frame Task

Although manual annotation of Situation Frames did not involve selecting a segment of text to justify each frame, LORELEI systems were required to output a single segment that justified both the frame type (e.g. *need for food*) and its place (e.g. *Hela*). These justification segments were subject to manual assessment for both the quality of the MT output and the utility of the selected segment for providing situational awareness within the context of the Situation Frame task. For all cases where the LORELEI system produced a Situation Frame that matched a gold standard reference Situation Frame on the same document, assessors reviewed that frame's justification segment across several dimensions. First, assessors were asked whether the MT for the selected justification segment was sufficiently intelligible to make subsequent assessment decisions, or if additional document context and/or the manual translation was required. The assessor then determined whether the situation frame type was justified by this segment, and if so, whether the place was also justified. If the segment was insufficient to justify either type or place, the assessor was shown the human translation for the segment and asked the same questions. A justification segment like "*People starving in Hela*" would be judged as being sufficiently intelligible and as justifying both the type and the place; while a justification segment like "*Food supply it run short*" would be assessed as intelligible and as justifying need but not place.

5.2 HyTER Annotation

To provide additional resources for diagnostic MT scoring in LORELEI, LDC produced a set of data annotated for HyTER. HyTER (Hybrid Translation Edit Rate) is an annotation approach that results in an exponential number of possible translations for a given sentence, thus producing large reference networks for translation evaluation (Dreyer and Marcu, 2012). We produced HyTER annotation for 645 English gold standard reference translation segments selected from the Uyghur Incident Language test set. One of the four available references was selected as the primary input reference for each segment. For each primary segment we performed two independent HyTER annotation passes, followed by a quality control pass on each segment. This effort yielded nearly 1.2×10^{15} unique meaning equivalents from the original 645 reference segments, with a median of 350,000 meaning equivalents per segment.

6 Maintaining Cross-Language Consistency

The research that underlies LORELEI system development relies in part on cross-lingual transfer approaches, as well as exploitation of language universals. As such, it is important for the RL language packs to be uniform and consistent in their design and implementation. At the same time, the RLs were selected specifically for their typological diversity. To achieve maximum compatibility across language packs while respecting the specific properties of each individual language, we adopted a number of strategies.

At the most basic level, the structure and core components for all language packs are the same, with consistent documentation and file formats across all corpora. All data collection and annotation efforts utilized a central database, enabling consistent handling of the data pipeline. We also used the same tools across all languages for data pre-processing wherever possible. For instance, all whitespace-delimited languages share a single tokenizer, whose rules were intentionally kept simple and were largely punctuation-based in order to increase uniformity across languages. While language-specific extensions to the rule set were possible, they were kept to the bare minimum.

We also used a shared inventory of tagsets and annotation labels across languages. For instance, Part of Speech tags in all RL lexicons are based on the 12 universal POS tags defined in Petrov et al. (2012), and all languages share the same entity and situation frame types and type definitions. Concepts shared across different annotation tasks utilized shared definitions and approaches. For instance, several LORELEI annotation tasks rely on annotators marking the extent of some phenomenon (like a named entity), so rules for selecting extents were defined in a uniform way across tasks and languages. All annotation is token-based rather than character-based; since tokens are defined using shared rules across all languages, this further reducing language-specific variation resulting from low-level annotation decisions.

To further enhance consistency in annotation, we developed template-based language-independent annotation guidelines which were then customized for each of the RL and IL languages as needed. We used the same policies across all languages for how to make decisions in the case of necessary language-specific extensions to the default approach. To achieve this, we first identified key questions about language features that could influence annotation, for instance, whether a language has possessive compound noun construction (e.g. Arabic idafa). Grammatical sketches for every language described whether the language possessed any of these annotation-relevant features, and if so how the phenomenon was realized in that language. The annotation guidelines template then provided a "menu" of options for how to localize the guidelines: if language has feature A, invoke section 3.6; if language has feature B, invoke section 4.8, and so on. This approach ensured consistency across languages within a task since languages with the same features get the same annotation treatment. It also ensured consistency in annotation approaches across tasks. Grammatical sketches themselves also follow the same template for all languages.

Finally, prior to data distribution, all language packs – both Representative and Incident – were subject to independent quality review by an external team including native speaker linguists. Among other factors, the independent QC team reviewed data for conformity to the pre-established language universal annotation policies and template-based guidelines.

7 Conclusion

Taken as a whole, the LORELEI Representative and Incident Language Packs represent a rich new resource for machine translation and natural language technology development in a low resource language setting. The Representative Language packs provide coverage of 23 typologically diverse languages, including some very low resource languages for which existing corpora are scarce. Beyond providing new data for these particular languages, the breadth of data and annotation types and the consistency of data components, corpus creation methods and annotation/translation approaches across the language packs is designed to support new research directions in the use of language universals and cross-language transfer. For MT research in particular the 23-way parallel text corpus represents a valuable new resource. The Incident Language Packs provide carefully curated test sets with gold standard translations and annotations for system development and testing. To date we have distributed representative language packs in 20 languages, as well as partial Representative Language Packs in three additional languages, to LORELEI performers. We have created seven Incident Language Packs; two additional Incident Language Packs are in progress to support the final LORELEI evaluation in 2019. LORELEI Representative and Incident Language Packs for all languages will begin to appear in the LDC Catalog in Fall 2019, making these resources broadly available to the research community at large.

8 Acknowledgements

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-15-C-0123. 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. The authors gratefully acknowledge Lori Levin, Chris Callison-Burch, Neville Ryant, Jonathan Wright, Brian Gainor, Christopher Caruso, Alex Shelmire, and the hundreds of LORELEI native speaker annotators for their contributions to this research.

References

Alvarez, Alison, Lori Levin, Robert Frederking, Simon Fung, Donna Gates, Jeff Good. 2006. The MILE Corpus for Less Commonly Taught Languages, In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers.* Association for Computational Linguistics, pages 5-8.

Arora, Sanjeev, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough to beat baseline for sentence embeddings. 2017. In International Conference on Learning Representations (ICLR 2017)

Compact Language Detector 2. [Online]. Available: https: //code.google.com/p/cld2/

DARPA LORELEI website, retrieved May 24, 2019. https://www.darpa.mil/program/low-resource-languages-for-emergent-incidents

Dreyer, Markus and Daniel Marcu. 2012. HyTER: Meaning-Equivalent Semantics for Translation Evaluation. *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Montreal, Canada 162–171.

Garland, Jennifer, Stephanie Strassel, Safa Ismael, Zhiyi Song, Haejoong Lee. 2012. Linguistic Resources for Genre-Independent Language Technologies: User-Generated Content in BOLT. *LREC 2012: 8th International Conference on Language Resources and Evaluation*, Istanbul, Turkey.

Kirov, Christo, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sebastian Mielke, Arya D. McCarthy, Sandra Kübler, David Yarowsky, Jason Eisner and Mans Hulden. 2018. UniMorph 2.0: Universal Morphology. *LREC 2018: 11th International Conference on Language Resources and Evaluation*, Miyazaki, Japan

Kiss, Tibor, Jan Strunkt. 2006. Unsupervised multilingual sentence boundary detection. *Computational Linguistics* 32: 455-525.

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

Kutuzov, Andrey, Mikhail Kopotev, Tatyana Sviridenko, Lyubov Ivanova. 2016. Clustering Comparable Corpora of Russian and Ukrainian Academic Texts: Word Embeddings and Semantic Fingerprints. *Proceedings of the Ninth Workshop on Building and Using Comparable Corpora,* Portorož, Slovenia

Ma, Xiaoyi. 2006. Champollion: A Robust Parallel Text Sentence Aligner. *LREC 2006: 5th International Conference on Language Resources and Evaluation*, Genoa, Italy

Ma, Xiaoyi, Mark Liberman. 1999. BITS: A Method for Bilingual Text Search over the Web. *Machine Translation Summit VII:* Singapore, September 13-17

Petrov, Slav, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. *LREC 2012: 8th International Conference on Language Resources and Evaluation*, Istanbul, Turkey.