CLARA: A New Generation of Researchers in Common Language Resources and Their Applications

Koenraad De Smedt[1], Erhard Hinrichs[2], Detmar Meurers[3], Inguna Skadiņa[4], Bolette Sandford Pedersen[5], Costanza Navarretta[6], Núria Bel[7], Krister Lindén[8], Markéta Lopatková[9], Jan Hajič[10], Gisle Andersen[11], Przemysław Lenkiewicz[12]

University of Bergen[1], University of Tübingen[2,3], Tilde[4], University of Copenhagen[5,6], University Pompeu Fabra[7], University of Helsinki[8], Charles University in Prague[9,10], Norwegian School of Economics[11], Max Planck Institute for Psycholinguistics[12]
Bergen, Norway[1,11], Tübingen, Germany[2,3], Riga, Latvia[4], Copenhagen, Denmark[5,6], Barcelona, Spain[7], Helsinki, Finland[8], Prague, Czech Republic[9,10], Nijmegen, The Netherlands[12]
desmedt@uib.no[1], erhard.hinrichs@uni-tuebingen.de[2], dm@sfs.uni-tuebingen.de[3], inguna.skadina@tilde.lv[4], bspedersen@hum.ku.dk[5], costanza@hum.ku.dk[6], nuria.bel@upf.edu[7], krister.linden@helsinki.fi[8], lopatkova@ufal.mff.cuni.cz[9], hajic@ufal.mff.cuni.cz[10], gisle.andersen@nhh.no[11], przemek.lenkiewicz@mpi.nl[12]

Abstract

CLARA (Common Language Resources and Their Applications) is a Marie Curie Initial Training Network which ran from 2009 until 2014 with the aim of providing researcher training in crucial areas related to language resources and infrastructure. The scope of the project was broad and included infrastructure design, lexical semantic modeling, domain modeling, multimedia and multimodal communication, applications, and parsing technologies and grammar models. An international consortium of 9 partners and 12 associate partners employed researchers in 19 new positions and organized a training program consisting of 10 thematic courses and summer/winter schools. The project has resulted in new theoretical insights as well as new resources and tools. Most importantly, the project has trained a new generation of researchers who can perform advanced research and development in language resources and technologies.

Keywords: researcher training, Marie Curie, CLARA

1. Background and Motivation

Qualified researchers are needed for the successful construction, curation, and application of new language resources and technologies. Even if highly qualified researchers had previously been produced at some institutions, our experience showed some bottlenecks. Recruiting for some new infrastructure positions showed a shortage of fully qualified candidates with knowledge of the state of the art. At the same time, there was also a shortage of funded PhD level positions at which researchers could gain research experience with state-of-the-art methods. Access to advanced courses and supervision was not evenly distributed among institutions and countries. There were insufficient incentives for mobility and cooperation across national borders in relevant researcher training. Finally, many of the research groups working with language resources in Europe are small and are unable to offer teaching and supervision in a wide area of current methods and techniques. In order to bring forth the next generation of qualified researchers, an advanced PhD level training project needed to be established to address these issues. The main aims have been to bundle the joint competencies from different institutions and countries, fund new positions for beginning researchers, promote their mobility across borders, and offer a training program far more comprehensive than what a single institution could produce.

A project called *Common Language Resources and Their Applications* (CLARA)[1] was therefore established as a Marie Curie Initial Training Network (ITN). The Marie Curie actions have been part of the People Specific programme in FP7, dedicated entirely to human resources in research, as officially stated in the workprogram: "To support the further development and consolidation of the European Research Area, the People Specific Programme's overall strategic objective is to make Europe more attractive for the best researchers." ITNs have been actions intended to provide researchers in the initial phase of their research career with advanced research competencies and skills. ITN projects have implied mobility, i.e. researchers have been employed in another country than their own, so as to promote transfer of knowledge across national borders.

Early Stage Researchers (ESRs) who do not yet have a doctoral degree have been the main targets of the action and have normally been enrolled in a doctoral study program, while Experienced Researchers (ERs) with a doctoral degree have been secondary targets. All researchers have formulated a Career Development Plan and the project has followed the European Charter of Researchers and Code of Conduct for the Recruitment of Researchers.[2]

2. Organization

The CLARA consortium has consisted of nine partners which have appointed and trained ESRs and ERs: University of Bergen (coordinator), University of Tübingen, Tilde (Riga), University of Copenhagen, University Pompeu Fabra (Barcelona), University of Helsinki, Charles

[1] Project website: http://clara.b.uib.no

[2] http://ec.europa.eu/euraxess/index.cfm/rights/whatIsAResearcher

University in Prague, Norwegian School of Economics (Bergen) and Max Planck Institute for Psycholinguistics (Nijmegen).

In addition, the following associate partners have contributed to training activities: Hungarian Academy of Sciences Research Institute for Linguistics, TEMIS (France), Evaluations and Language Resources Distribution Agency (ELDA, France), University of Zagreb, Uni Research (Norway), Lexical Computing (UK), Universität Wien, Ilisimatusarfik/University of Greenland, Università degli Studi di Roma "La Sapienza", Universität des Saarlandes, Infomedia (Denmark) and RWTH Aachen University.

The CLARA project started on December 1, 2009 and ran until March 31, 2014 with the aim of training researchers working towards the next generation of language resources, data-intensive language models and applications. This paper highlights its main activities and achievements.

The CLARA project has filled 17 ESR fellowships[3] and two ER fellowships. Positions were announced for periods ranging between 21 and 36 months, and were published on Euraxess, on the project website and other channels. Selection of candidates was performed by staff members, taking into account eligibility criteria, qualifications, equal opportunities and project goals. Successful candidates came mostly from European countries, but also from Africa, Asia and North America. There was an optimal gender balance. The project has also had four visiting researchers.

3. Subprojects and results

Researchers have been employed at one of the project partner organizations, where they joined in a subproject, as described below. These subprojects have been chosen to address existing research opportunities at the partner organizations, and where the necessary supervision capacity was present. In many cases the researchers have produced resources and/or tools which are distributed or deployed in the wider research community.

3.1. Designing and Testing Common Infrastructures

In this subproject at the Max Planck Institute for Psycholinguistics, the Language Archive[4] and its software components have been extended.

In his ESR position at this institute, *Binyam Gebrekidan Gebre* has developed machine-learning systems to speed up the annotation of multimedia in general, and human gestures in particular. He has developed innovative methods for speaker/signer diarization, i.e. partitioning an input video stream into segments according to speaker identity. This task is very important for search and retrieval of information in digitized dialogs. Based on the hypothesis that the gesturer is the speaker (Gebre et al., 2013c; Gebre et al., 2013b), he has developed an efficient algorithm for detecting gestures and attributing them to speakers, based on motion history images (Gebre et al., 2014). He has also worked on the identification of sign languages from video (Gebre et al., 2013a), native languages from text (Gebre et al., 2013d) and language varieties from text (Zampieri and Gebre, 2012; Zampieri et al., 2012). The importance of language identification is well recognized. The source code has been made freely available for public use.[5] The Language Archive has added new digital media from researchers worldwide and is connected to the Virtual Language Observatory in cooperation with CLARIN.

Another ESR appointed at this institute, *Anna Lenkiewicz*, has developed a system supporting knowledge discovery in linguistic recordings (Lenkiewicz et al., 2012a; Lenkiewicz and Drude, 2013). A domain-specific query language allows searching for specific patterns in recordings using rule-based and machine-learning approaches. The main premise for this work has been to decrease the time of manual annotation and to enable interoperability between team members involved in the annotation process. Anna Lenkiewicz has defined a media query language and interface, which allows searching for specific patterns or features in audio and video recordings, which greatly speeds up annotation (Lenkiewicz et al., 2012a; Lenkiewicz et al., 2011). The idea for this topic was inspired by the need for automatic annotation tools which will enable researchers to efficiently annotate hours of recordings with very little supervision. Furthermore, pattern recognition and search components have been integrated in a multimodal annotation framework.

3.2. Lexical Semantic Modeling

Lexical and conceptual resources have been the topic of work in this subproject.

Carla Parra, an ESR at the University of Bergen, has compiled, encoded and aligned a bilingual German/Spanish corpus which is explored for translations of German compounds into and from Spanish phrases (Parra Escartín, 2012; Parra Escartín, 2013). The corpus, called TRIS and publicly available via META-SHARE, comprises translated documents from the DG Enterprise and Industry Project of the EC, from which documents in three technical domains have been selected. Carla Parra spent a three month secondment at the Human Language Technology and Pattern Recognition Group of the RWTH University Aachen, where she ran several SMT experiments with their Jane system. The TRIS corpus was used in different experimental set-ups devised to test to which extent compound preprocessing and other strategies have an impact in SMT tasks from German into Spanish, a language pair not very much researched so far in SMT (Parra Escartín et al., 2014). Evaluation of the output of these experiments has revealed that different compound splitters have a different impact in machine translation quality (Parra Escartín, 2014).

At the University of Copenhagen, corpora were annotated with respect to regular polysemy in English, Spanish and Danish by ESR *Héctor Martínez Alonso*. He has worked on semi-supervised recognition of metonymic senses for regular polysemy and has researched strategies to capture literal-metonymic senses using a continous representation instead of discrete categories (Martínez Alonso et al., 2012; Martínez Alonso et al., 2011; Martínez Alonso et al., 2013; Romeo et al., 2013). The study includes cross-lingual aspects of semantic corpus annotation in cooperation with the

[3] One of the positions was filled but did not book significant results due to the researcher taking leave early.

[4] http://tla.mpi.nl

[5] https://bitbucket.org/binyam

University Pompeu Fabra.[6] Few PhD students at Humanities faculties go spend some time at an industrial company, but that is precisely what Héctor Martínez has done. He spent a secondment at Infomedia, a Danish IT company which has large text databases containing millions of nouns which were useful for his research.

At the University Pompeu Fabra, ESR *Silvia Necşulescu* developed a novel method for the classification of semantic relations. Whereas the recall of current systems is upper bound limited by the co-occurrences of word pairs in the same context, her contribution is based on a graph representation of the corpus where information from different sentences is gathered. By mixing information across sentence boundaries, more information is leveraged for the classification, so that recall is improved (Necşulescu, 2011). She developed graph unification-based techniques to automatically merge lexical resources (Necşulescu et al., 2011; Bel et al., 2011; Padró et al., 2011). This method is promising for the future cost-effective development of richer resources. She also spent a secondment at Temis GmbH Heidelberg, where she worked on the JobimTex system.

3.3. Next Generation Domain Modeling

The goals of this subproject have been the construction, harmonization and management of multilingual terminological resources.

Pedro Patiño, an ESR at the Norwegian School of Economics, has compiled a specialized parallel corpus of 16 English and Spanish Free Trade Agreements, aligned it at sentence level and studied it with respect to specialized collocations found in legal and economics domains (Mira and Patiño, forthcoming; Patiño García, 2013; Patiño, 2013; Patiño, 2011). The corpus contains around 1.37 million words in the English section and 1.48 million words in its Spanish counterpart, plus 60,000 words each in the Spanish-Norwegian and English-Norwegian subcorpus. Term lists were extracted using collocation patterns and validated as terms by consulting specialist dictionaries of the field of economics, finance and international trade. Such terminology lists may contribute to enrich term bases and electronic dictionaries to support human and machine translation.

Anne Schumann, an ESR appointed at Tilde and a PhD student of the University of Vienna, has worked on automated methods for context extraction and enrichment of a multilingual terminology database with knowledge-rich contexts. She has performed a feasibility study on pattern-based knowledge-rich context extraction in Russian for the automotive domain (Schumann, 2011b; Schumann, 2011c) and a bilingual study of knowledge-rich context extraction in Russian and German (Schumann, 2011a). She has also developed a pipeline for extraction of Russian and German knowledge-rich context candidates from text corpora and a ranking algorithm that supports the selection of high-quality context (Schumann, 2012a; Schumann, 2012b; Schumann, 2012c). One of the resulting available resources is a German and Russian gold standard for evaluation of knowledge-rich context extraction methods.

[6]The sense-annotated corpus is at http://metashare.cst.dk/repository/search/?q=regular+polysemy.

3.4. Multimedia and Multimodal Communication Modeling

In this subproject, the Max Planck Institute and the University of Copenhagen have cooperated on innovative methods and emerging standards for audio and video annotation.

Przemysław Lenkiewicz, appointed as ER at the Max Planck Institute, has collaborated with two of the ESRs on their work on multimodal annotation and analysis of linguistic resources (Gebre et al., 2012; Lenkiewicz et al., 2012a). His research work has been focused on the audio and video processing algorithms for automated annotation of linguistic recordings (Lenkiewicz et al., 2012b; Lenkiewicz et al., 2011) and the human-computer interaction aspect of the developed tools (Lenkiewicz et al., 2012c; Wittenburg et al., 2012). He has also collaborated with partners from the University of Copenhagen on the evaluation of algorithms developed at the MPI.

ESR *Magdalena Lis* at the University of Copenhagen has contributed to the collection, annotation and analysis of multimodal corpora (Lis, 2012c; Lis, 2012a). She has studied and formalized the semantic relations between speech and iconic hand gestures in narrative and conversational multimodal corpora in more languages using Wordnet (Lis, 2012b). Classifiers were trained on the annotations to predict types of gesture form from the semantics of co-speech (Lis and Navarretta, forthcoming; Lis and Parrill, forthcoming). The results of this work are promising from both a cognitive and applied point of view. Similarities and differences in multimodal feedback in spontaneous multimodal conversations in Danish and Polish (Navarretta and Lis, 2013) have been investigated. A classifier has been trained on the multimodal annotations in one language to annotate feedback with head movements in the other language (Navarretta and Lis, 2014).

3.5. Applications

In this subproject, the Latvian language technology company Tilde and the Charles University in Prague have cooperated on methods that facilitate the development of missing translation tools and relevant resources necessary for under-resourced languages. A second LT application, Computer Assisted Language Learning, has been researched at the University of Tübingen.

Septina Dian Larasati, an ESR at Tilde and PhD student in Prague, has worked on translation tools and resources for under-resourced languages. Her main focus has been on the Indonesian language. She has investigated different machine translation strategies to find the most suitable methods for Indonesian as under-resourced language. Rule-based machine translation was applied for translation between Indonesian and Malaysian (Susanto et al., 2012). On the other hand, statistical machine translation was applied to the Indonesian/English translation task, for which she investigated methods and algorithms that improve SMT training for under-resourced languages (Larasati, 2012a; Larasati, 2012c; Larasati, 2012d). Septina Dian Larasati has also produced several language resources for Indonesian, including a Indonesian dependency treebank (Green et al., 2012a) and an Indonesian-English parallel corpus (Larasati, 2012b).

Loganathan Ramasamy, appointed as ESR at Charles Uni-

versity in Prague, has developed effective methods to create resources for data-driven language technologies for Tamil, an under-resourced language. He has developed tools and prepare a dependency-based treebank for Tamil (Ramasamy and Žabokrtský, 2012; Ramasamy and Žabokrtský, 2011a; Ramasamy and Žabokrtský, 2011b) and harvested English-Tamil parallel corpus (Ramasamy et al., 2012a) from the web. Both the treebank[7] and the parallel corpus[8] have been released for public use and are freely available for research. At present, the English-Tamil parallel corpus is the largest publicly available one for Tamil and is being used by machine translation researchers in the field. He has also explored methodologies (Ramasamy et al., 2012b; Ramasamy et al., 2012a; Green et al., 2012b) suitable for typologically different languages (such as agglutinative languages) that are also under-resourced.

Sowmya Vajjala has worked as an ESR in Tübingen at the interface of NLP and language learning. Her research has resulted in the creation of a comprehensive corpus of graded texts for training readability models (Vajjala and Meurers, 2012) and a rich set of linguistic features for building a robust readability model. This feature set was shown to be successful for a broad range of materials, such as web documents (Vajjala and Meurers, 2013), standard graded texts (Vajjala and Meurers, 2014c), spoken language transcripts (Vajjala and Meurers, 2014a), and for text simplification evaluation (Vajjala and Meurers, 2014b). The approach is currently the best non-commercial readability model for English, and the second best including commercial systems. A web interface to this reading level assessment approach is currently under development. The methods developed for English were shown to be equally successful for assessing and classifying German texts (Hancke et al., 2012). In the broader CLARA context, Sowmya Vajjala has also collaborated with researchers in Prague to study morphological segmentation in agglutinative languages (Ramasamy et al., 2012b).

3.6. Parsing Technologies and Grammar Models

This subproject has aimed at novel approaches to parsing ranging from shallow to deep parsing models.

Mans Hulden has been employed as an ER at the University of Helsinki, where he has developed new methods for efficiently implementing constraint grammar rules with finite state technologies (Hulden, 2011a; Hulden, 2011b), learning phonological replacement rules from parallel corpora (Hulden et al., 2011; Uria et al., 2011), and converting context-free grammars and probabilistic context-free grammars into parallel FSMs for parsing (Hulden, 2011c). The research group at the University of Helsinki has released the open source toolset HFST.[9]

Senka Drobac, employed as ESR at the same institution, has developed a reduction of Xerox XFST operations to generalized restriction and other basic finite-state operations, providing a framework for any finite-state library as a driver for XFST tools (Lindén et al., 2013b). She has also developed methods for hyperminimization of morphological lexicons, providing a way to minimize lexicons and grammars beyond their minimal FSM size (Lindén et al., 2013a). This is relevant for synthetic languages such as Greenlandic which have an extremely high number of possible word forms. The Greenlandic lexicon can be hyper-minimized with a size reduction of approximately 90 % without loss of information and with only a 10 % reduction in look-up speed (Drobac et al., 2014). Other languages can also profit from this method, but the memory size reduction is less pronounced. This brings the finite-state Greenlandic lexicon into the same size-range as that of other languages allowing practical applications such as spell-checking on PCs and mobile devices.

At the University of Bergen, ESR *Bamba Dione* has implemented a tokenizer and morphological component for Wolof, and under-resourced language (Dione, 2012b). He has also implemented a deep grammar for Wolof in the LFG formalism, using the XLE parser and the INESS parsebanking infrastructure. His grammar provides linguistically well motivated analyses of challenging constructions in Wolof including clitics, clefts, valency change and complex predicates (Dione, 2013a; Dione, 2012a; Dione, 2013b), in cooperation with the ParGram network project (Sulger et al., 2013). He has also performed extensive experiments to manage ambiguity which is pervasive in Wolof. Different avenues were explored to this end, including the formal encoding of noun class indeterminacy, lexical specifications, the use of a probabilistic and a Constraint Grammar to prune the search space (Dione, 2014), and optimality marks. The parsing system is further controlled by packing ambiguities and discriminant-based techniques for parse disambiguation.

At the Charles University in Prague, ESR *Nathan Green* has studied the influence of noun phrase bracketing—as well as of the overall dependency annotation scheme—on dependency parsing and statistical machine translation. While the effect of various NP bracketing schemes for standard parsing has been minimal, it does help machine translation (Green, 2011b; Green and Žabokrtský, 2012a; Popel et al., 2011). Effects of various head/dependency annotation standards on various NLP tasks were also studied (Green and Žabokrtský, 2012b; Green, 2011a). Nathan Green has proved that ensemble parsing techniques have an influence on syntax-based machine translation both in manual and automatic evaluation. A stronger correlation between parser accuracy and the NIST rather than the more commonly used BLEU metric was shown (Green and Žabokrtský, 2013). In cooperation with other researchers, new avenues in parsing were explored and tested on Tamil (Green et al., 2012b).

Corina Dima, an ESR at the University of Tübingen, has focused on creating a natural language search facility for querying large metadata repositories. The initial work concentrated on building a question treebank (Dima and Hinrichs, 2011) that could serve both as training material for a statistical parser and as genuine user input for the final system. Then the focus turned to querying Linked Data repositories: she built a question answering system (Dima, 2013) that is able to convert a natural language question to

[7] http://ufal.mff.cuni.cz/~ramasamy/tamiltb/0.1, also available on http://clarino.uib.no/iness

[8] http://ufal.mff.cuni.cz/~ramasamy/parallel/html

[9] http://hfst.sf.net/

its Linked Data representation. The system maps the syntactic patterns in a question to their most likely RDF representation, while constructing and keeping track of multiple possible interpretations. It was tested on general encyclopaedic questions whose answers could be found in DBpedia, a large, multi-lingual knowledge base extracted from Wikipedia in the context of the QALD3 competition at CLEF 2013.

Jianqiang Ma has been employed as an ESR at the University of Tübingen. In the context of practical NLP tasks such as Chinese word segmentation, he has developed computational models of Chinese word formation and algorithmic approaches to Chinese word structure annotation. He had generalized word-based models for Chinese word segmentation to a phrase-based one (Ma et al., 2012). Having shown limitations of the traditional sequence labeling framework for word segmentation (Ma and Gerdemann, 2013), he proposed word-structure parsing models, for which he developed a semi-automatic word structure annotation method (Ma et al., 2012). He has also produced new, syntactically inspired algorithms to automatically refine the word structure annotation for more accurate parsing (Ma, 2014). The refined annotation is publicly available.[10] Such annotated word structures can benefit a variety of Chinese NLP tasks, such as word segmentation, POS tagging and syntactic parsing. The word structure parsing work is expected to attract more research efforts from others to this new direction of Chinese NLP.

4. Europe-wide training program

While the CLARA project has contributed to significant research results, as described in the previous section, its primary goal has been to offer advanced researcher training in language resources and their applications. ESRs were locally enrolled in PhD programs and were assigned a supervisor. Most appointed researchers spent secondments at other partners or associate partners in the project, often in a different country, and sometimes at companies as well as academic institutions. These secondments have given them access to supplementary expertise, data and tools.

Furthermore, the consortium offered a Joint Training Programme of thematic courses covering important scientific areas in the project. The aim was to bundle teaching competencies in ways that go beyond what individual organizations were able to offer. The project organized the following specific training events, which attempted to cover the needs of several researchers in the project. These events, most of which had hands-on sessions, were also open to researchers from outside the project.

1. Summer School in Advanced Resource Creation, Archiving and Usage (Nijmegen, 5-16 July 2010): training in the creation of language resources from multimedia streams, archiving the resulting complex resource types, access and analysis via state-of-the-art web applications for their enrichment and virtual collections.
2. Thematic Training Course on Methods and Technologies for Consolidating and Harmonising Terminological Resources (Bergen, 13-17 Sep. 2010): state-of-the-art terminology work, including multilingual perspectives, terminological variation, corpus-based term extraction, ontology-based domain recognition, and the consolidation and integration of existing terminological resources.
3. Thematic Training Course on Methods and Technologies for Consolidating and Harmonising Treebank Annotation (Prague, 13-17 Dec. 2010): construction and exploitation of treebanks in constituency, dependency and LFG frameworks, and state-of-the-art syntactic search tools.
4. Thematic Training Course on Processing Morphologically Rich Languages (Budapest, 11-15 April 2011): morphological analysis and lemmatization with state-of-the-art tools, including HFST; especially relevant for languages like Turkish, Finnish, Hungarian, and Sámi which have relatively free word order and highly productive morphology.
5. Thematic Training Course on Multilingual Lexical Resources and Tools (Bergen, 20-23 June 2011): lexical relations in computational lexicons, wordnets and thesauri, with methods for their construction and exploitation; standardization of lexical resources in LMF (Lexical Mapping Framework) and demonstrations of SketchEngine and Semantic Mirrors tools.
6. Summer School on Infrastructure Tool Development (Nijmegen, 5-12 July 2011): audio, video and text processing for automated content analysis and automated annotation creation with advanced tools including ELAN plugins.
7. Summer School in Semantic and Nonverbal Corpus Annotation and Evaluation (Copenhagen, 15-26 Aug. 2011): semantic corpus annotation (sense ambiguity, semantic roles) and multimodal (verbal and non-verbal) annotation of video-recorded interactions; tools such as Stamp, Cornerstone, Jubilee, GATE, and ANVIL.
8. Industrial Career Training Course on Product Planning for Next Generation Information Access Technology Solutions (Dubrovnik, 20-23 Sep. 2011): complementary skills such as industrial project and product planning, CV presentation, and interviews, which are often overlooked in traditional academic curricula.
9. Winter School on New Developments in Computational Linguistics (Prague, 13-17 Feb. 2012): innovative approaches to machine translation, parsing and the use of annotated language resources.
10. Thematic Training Course on Evaluation of Human Language Technologies (Paris, Nov. 2012): awareness of the important role of evaluation and described the role of automatic vs. human evaluation; evaluation metrics and testing procedures for several tasks, also targeting speech and multimodal technologies.

[10] http://www.sfs.uni-tuebingen.de/~jma/word_str.txt

In addition, CLARA participated actively at the CHAT 2011 and CHAT 2012 workshops on the creation, harmonization and application of terminology resources.

Training events and workshops have played an important role in work package integration and provided a meeting point for the project partners and external participants. They have also facilitated the initiation of concrete research cooperations, as witnessed by joint publications. Visiting researchers Jens Edlund, Susan Brown, Rita Temmerman and Sandra Kübler have further contributed to the training.

5. Conclusion

The present paper is the first publicly presented overall report on the CLARA project, which has successfully concluded its planned training program. The network has been effective in exposing a broad slate of young people to the state of the art and supporting them in performing and publishing advanced research. More than 100 publications resulting from the project describe the research data, methods and tools in detail.

Several languages were targeted with a wide range of methods for their analysis, annotation and modeling. Many of the fellows had a background in under-resourced languages and their work has contributed to new resources as well as new methods to overcome the data scarcity for such languages (including Greenlandic, Indonesian, Malaysian, Tamil, Wolof, and sign languages). The project has paid attention to metadata standards, cataloguing, search and IPR. Results in the form of presentations and publications, as well as new language resources and tools, have been disseminated through academic publication channels, through participation at scientific meetings (such as LREC), through repositories at META-SHARE and CLARIN, on the CLARA website and through contact with other projects (e.g. TTC, INESS and ParGram).

The main impact of the project has been twofold. On the one hand, this collective work, taken as a whole, has been a major contribution to the creation of language research infrastructures such as META-SHARE and CLARIN, which are reusing and distributing many of the resources, tools and strategies developed in CLARA.

On the other hand, through mobility across borders and with the benefit of international courses, a transfer of competency has been achieved, including joint PhD degrees, which certify recognized qualifications in two countries and widen career perspectives. The involvement of industrial partners and associate partners who hired early stage researchers or hosted them for secondments, and contributed to training courses, has given early stage researchers a deep sense of the requirements and potential of an industrial career in language technology.

In conclusion, a new generation of researchers has started to contribute to high quality language resources, tools, infrastructures, applications and their evaluation.

6. Acknowledgements

The research reported in this paper has received support from the European Union through the Marie Curie Actions, grant 238405. The authors acknowledge the contributions of the CLARA fellows to this text.

7. References

Bel, Núria, Padró, Muntsa, and Necşulescu, Silvia. (2011). A method towards the fully automatic merging of lexical resources. In *Proceedings of the Workshop on Language Resources, Technology and Services in the Sharing Paradigm*, pages 8–15, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Dima, Corina and Hinrichs, Erhard. (2011). A semi-automatic, iterative method for creating a domain-specific treebank. In *Proceedings of the 8th International Conference on Recent Advances in Natural Language Processing*, pages 413–419, Hissar, Bulgaria.

Dima, Corina. (2013). Intui2: A prototype system for question answering over linked data. In Forner, Pamela, Navigli, Roberto, and Tufis, Dan, editors, *CLEF 2013 Evaluation Labs and Workshop, Online Working Notes*, Valencia, Spain.

Dione, Cheikh M. Bamba. (2012a). An LFG approach to Wolof cleft constructions. In *Proceedings of the LFG'12 Conference*, LFG Online Proceedings, pages 157–176, Stanford, CA. CSLI Publications.

Dione, Cheikh M. Bamba. (2012b). A morphological analyzer for Wolof using finite-state techniques. In *Proceedings of the 8th LREC*, pages 894–901, Istanbul, Turkey. ELRA.

Dione, Cheikh M. Bamba. (2013a). Handling Wolof clitics in LFG. In Salvesen, Christine Meklenborg and Helland, Hans Petter, editors, *Challenging Clitics*, pages 87–118. John Benjamins Publishing Company, Amsterdam.

Dione, Cheikh M. Bamba. (2013b). Valency change and complex predicates in Wolof: An LFG account. In Butt, Miriam and King, Tracy Holloway, editors, *Proceedings of the LFG '13 Conference*, Stanford, CA. CSLI Publ.

Dione, Cheikh M. Bamba. (2014). Pruning the search space of the Wolof LFG grammar using a probabilistic and a constraint grammar parser. In *Proceedings of the 9th LREC*, Reykjavik, Iceland. ELRA.

Drobac, Senka, Lindén, Krister, Pirinen, Tommi A., and Silfverberg, Miikka. (2014). Heuristic hyper-minimization of finite state lexicons. In *Proceedings of the 9th LREC*, Reykjavik, Iceland. ELRA.

Gebre, Binyam Gebrekidan, Wittenburg, Peter, and Lenkiewicz, Przemyslaw. (2012). Towards automatic gesture stroke detection. In *Proceedings of the 8th LREC*, Istanbul, Turkey. ELRA.

Gebre, Binyam Gebrekidan, Wittenburg, Peter, and Heskes, Tom. (2013a). Automatic sign language identification. In *Proceedings of ICIP 2013*, pages 2626–2630.

Gebre, Binyam Gebrekidan, Wittenburg, Peter, and Heskes, Tom. (2013b). Automatic signer diarization – the mover is the signer approach. In *Proceedings of the CVPR Workshop on Language for Vision*.

Gebre, Binyam Gebrekidan, Wittenburg, Peter, and Heskes, Tom. (2013c). The gesturer is the speaker. In *Proceedings of ICASSP 2013*.

Gebre, Binyam Gebrekidan, Zampieri, Marcos, Wittenburg, Peter, and Heskes, Tom. (2013d). Improving native language identification with tf-idf weighting. In *Proceed-*

ings of the 8th NAACL Workshop on Innovative Use of NLP for Building Educational Applications.

Gebre, Binyam Gebrekidan, Wittenburg, Peter, Heskes, Tom, and Drude, Sebastian. (2014). Motion history images for online speaker/signer diarization. In *Proceedings of the 2014 IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE.

Green, Nathan and Žabokrtský, Zdeněk. (2012a). Ensemble parsing and its effect on machine translation. Technical Report 48, Charles University, Prague, Czech Republic.

Green, Nathan and Žabokrtský, Zdeněk. (2012b). Hybrid combination of constituency and dependency trees into an Ensemble Dependency Parser. In *Workshop on Innovative Hybrid Approaches to the Processing of Textual Data*, pages 19–26, Avignon, France. ACL.

Green, Nathan and Žabokrtský, Zdeněk. (2013). Improvements to syntax-based machine translation using ensemble dependency parsers. In *Proceedings of the 2nd Workshop on Hybrid Approaches to Translation at ACL*, pages 19–24, Sofia, Bulgaria. Bălgarska akademija na naukite, ACL.

Green, Nathan, Larasati, Septina Dian, and Žabokrtský, Zdeněk. (2012a). Indonesian dependency treebank: Annotation and parsing. In *Proceedings of the 26th PACLIC*, pages 137–145, Bali, Indonesia. Faculty of Computer Science, Universitas Indonesia.

Green, Nathan, Ramasamy, Loganathan, and Žabokrtský, Zdeněk. (2012b). Using an SVM Ensemble System for improved Tamil dependency parsing. In *Proceedings of the ACL Joint Workshop on Statistical Parsing and Semantic Processing of Morphologically Rich Languages*, pages 72–77, Jeju, Republic of Korea. ACL.

Green, Nathan. (2011a). Dependency parsing. In *WDS 2011 Proceedings of Contributed Papers*, pages 137–142, Prague, Czech Republic.

Green, Nathan. (2011b). Effects of noun phrase bracketing in dependency parsing and machine translation. In *Proceedings of the ACL 2011 Student Session*, pages 69–74, Portland, OR, USA. ACL.

Hancke, Julia, Vajjala, Sowmya, and Meurers, Detmar. (2012). Readability classification for German using lexical, syntactic, and morphological features. In *Proceedings of the 24th COLING*, volume Technical Papers, pages 1063–1080, Mumbai, India.

Hulden, Mans, Alegria, Iñaki, Etxeberria, Izaskun, and Maritxalar, Montse. (2011). Learning word-level dialectal variation as phonological replacement rules using a limited parallel corpus. In *Proceedings of the 1st Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*, pages 39–48, Edinburgh, Scotland. ACL.

Hulden, Mans. (2011a). Constraint grammar parsing with left and right sequential finite transducers. In *Proceedings of the International Workshop on Finite State Methods in Natural Language Processing*.

Hulden, Mans. (2011b). Finite-State Technology. In Mitkov, Ruslan, editor, *The Oxford Handbook of Computational Linguistics*. Oxford University Press, 2nd edition.

Hulden, Mans. (2011c). Parsing CFGs and PCFGs with a Chomsky-Schützenberger representation. In *Human Language Technology. Challenges for Computer Science and Linguistics*, volume 6562 of *Lecture Notes in Artificial Intelligence*, pages 151–160. Springer.

Larasati, Septina Dian. (2012a). Handling Indonesian clitics: A dataset comparison for an Indonesian-English statistical machine translation system. In *Proceedings of the 26th PACLIC*, pages 146–152, Bali, Indonesia. Faculty of Computer Science, Universitas Indonesia.

Larasati, Septina Dian. (2012b). IDENTIC corpus: Morphologically enriched Indonesian-English parallel corpus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. ELRA.

Larasati, Septina Dian. (2012c). Improving word alignment by exploiting adapted word similarity. In *Proceedings of the Workshop on Monolingual Machine Translation at AMTA 2012*, pages 41–45, San Diego, CA. The Association for Machine Translation in the Americas.

Larasati, Septina Dian. (2012d). Towards an Indonesian-English SMT system: A case study of an under-studied and under-resourced language, Indonesian. In *WDS'12 Proceedings of Contributed Papers*, volume I, pages 123–129, Prague, Czech Republic. MatfyzPress.

Lenkiewicz, Anna and Drude, Sebastian. (2013). Automatic annotation of linguistic 2D and Kinect recordings with the Media Query Language for Elan. In *Proceedings of the Digital Humanities Conference*.

Lenkiewicz, Przemyslaw, Wittenburg, Peter, Gebre, Binyam Gebrekidan, Lenkiewicz, Anna, Schreer, Oliver, and Masneri, Stefano. (2011). Application of video processing methods for linguistic research. In *Human language technologies as a challenge for computer science and linguistics. Proceedings of the 5th Language and Technology Conference*, pages 561–564, Poznań, Poland. Adam Mickiewicz University.

Lenkiewicz, Anna, Lis, Magdalena, and Lenkiewicz, Przemyslaw. (2012a). Linguistic concepts described with Media Query Language for automated annotation. In *Proceedings of the Digital Humanities Conference*, pages 477–479, Hamburg.

Lenkiewicz, Przemyslaw, Gebre, Binyam Gebrekidan, Schreer, Oliver, Masneri, Stefano, Schneider, Daniel, and Tschöpel, Sebastian. (2012b). AVATecH – automated annotation through audio and video analysis. In *Proceedings of the 8th LREC*, Istanbul, Turkey. ELRA.

Lenkiewicz, Przemyslaw, Van Uytvanck, Dieter, Wittenburg, Peter, and Drude, Sebastian. (2012c). Towards automated annotation of audio and video recordings by application of advanced web-services. In *13th INTERSPEECH*, Portland, OR, USA.

Lindén, Krister, Axelson, Erik, Drobac, Senka, Hardwick, Sam, Kuokkala, Juha, Niemi, Jyrki, Pirinen, Tommi, and Silfverberg, Miikka. (2013a). HFST – a system for creating NLP tools. In Mahlow, Cerstin and Piotrowski, Michael, editors, *Systems and Frameworks for Computational Morphology*, volume 380 of *Communica-*

tions in Computer and Information Science, pages 53–71. Springer Verlag.

Lindén, Krister, Axelson, Erik, Drobac, Senka, Hardwick, Sam, Silfverberg, Miikka, and Pirinen, Tommi. (2013b). Using HFST for creating computational linguistic applications. In Przepiórkowski, Adam, Piasecki, Maciej, Jassem, Krzysztof, and Fuglewicz, Piotr, editors, *Computational Linguistics – Applications*, volume 458 of *Studies in Computational Intelligence*, pages 3–25. Springer Verlag.

Lis, Magdalena and Navarretta, Costanza. (forthcoming). Classifying the form of iconic hand gestures from the linguistic categorization of co-occurring verbs. In *Proceedings of the 1st European Symposium on Multimodal Communication*, Valletta, Malta.

Lis, Magdalena and Parrill, Fey. (forthcoming). Referent type and its verbal and gestural representation: A test on English multimodal corpus and wordnet3.1. In *Proceedings of the 1st European Symposium on Multimodal Communication*, Valletta, Malta.

Lis, Magdalena. (2012a). Annotation scheme for multimodal communication: Employing plWordNet 1.5. In *Proceedings of the Formal and Computational Approaches to Multimodal Communication Workshop. 24th ESSLLI*, Opole, Poland.

Lis, Magdalena. (2012b). Influencing gestural representation of eventualities: insights from ontology. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction*, pages 281–288, New York, NY. ACM.

Lis, Magdalena. (2012c). Polish multimodal corpus – a collection of referential gestures. In *Proceedings of the 8th LREC*, Istanbul, Turkey. ELRA.

Ma, Jianqiang and Gerdemann, Dale. (2013). Distributional evidence and beyond: the success and limitations of machine learning in Chinese word segmentation. *Research in Computing Science*, 70:17–30.

Ma, Jianqiang, Kit, Chunyu, and Gerdemann, Dale. (2012). Semi-automatic annotation of Chinese word structure. In *Proceedings of the 2nd CIPS-SIGHAN Joint Conference on Chinese Language Processing*, pages 9–17, Tianjin, China. ACL.

Ma, Jianqiang. (2014). Automatic annotation of syntactic categories in chinese word structures. In *Proceedings of the 9th LREC*, Reykjavik, Iceland.

Martínez Alonso, Héctor, Bel, Núria, and Pedersen, Bolette Sandford. (2011). Identification of sense selection in regular polysemy using shallow features. In *Proceedings of the 18th NoDaLiDa*, number 11 in NEALT Proceedings Series, pages 18–25.

Martínez Alonso, Héctor, Bel, Núria, and Pedersen, Bolette Sandford. (2012). A voting scheme to detect semantic underspecification. In *Proceedings of the 8th LREC*, Istanbul, Turkey. ELRA.

Martínez Alonso, Héctor, Sandford Pedersen, Bolette, and Bel, Núria. (2013). Annotation of regular polysemy and underspecification. In *Proceedings of the 51st ACL*, pages 725–730, Sofia, Bulgaria. ACL.

Mira, Germán and Patiño, Pedro. (forthcoming). La traducción de la fraseología especializada. In Colín, Marisela, editor, *Manual de traducción de textos especializados. Nuevos enfoques, nuevas metodologías*, pages 20–35. UNAM, México.

Navarretta, Costanza and Lis, Magdalena. (2013). Multimodal feedback expressions in Danish and Polish feedback expressions. In *Proceedings of the 4th Nordic Symposium on Multimodal Communication*, pages 55–62, Linköping, Sweden. NEALT, LiU Electronic Press.

Navarretta, Costanza and Lis, Magdalena. (2014). Transfer learning of feedback head expressions in comparable multimodal danish and polish corpora. In *Proceedings of the 9th LREC*, Reykjavik, Iceland. ELRA.

Necşulescu, Silvia, Bel, Núria, Padró, Muntsa, Marimon, Montserrat, and Revilla, Eva. (2011). Towards the automatic merging of language resources. In *Proceedings of the International Workshop on Lexical Resources*, Ljubljana, Slovenia.

Necşulescu, Silvia. (2011). Automatic acquisition of possible contexts for low frequent words. In *Proceedings of the Student Research Workshop at RANLP*, pages 121–126, Hissar, Bulgaria.

Padró, Muntsa, Bel, Núria, and Necşulescu, Silvia. (2011). Towards the automatic merging of lexical resources: Automatic mapping. In *Proceedings of the 8th International Conference on Recent Advances in Natural Language Processing*, pages 296–301, Hissar, Bulgaria.

Parra Escartín, Carla, Peitz, Stephan, and Ney, Hermann. (2014). German compounds and statistical machine translation. can they get along? In *Proceedings of the 10th Workshop on Multiword Expressions*, Gothenburg, Sweden.

Parra Escartín, Carla. (2012). Design and compilation of a specialized Spanish-German parallel corpus. In *Proceedings of the 8th LREC*, pages 2199–2206, Istanbul, Turkey. ELRA.

Parra Escartín, Carla. (2013). Encoding a parallel corpus: The TRIS corpus experience. *Bergen Language and Linguistics Studies*, 3(1):61–78.

Parra Escartín, Carla. (2014). Chasing the perfect splitter: A comparison of different compound splitting tools. In *Proceedings of the 9th LREC*, Reykjavik, Iceland. ELRA.

Patiño García, Pedro. (2013). FTA corpus: a parallel corpus of English and Spanish Free Trade Agreements for the study of specialized collocations. *Bergen Language and Linguistics Studies*, 3(1):81–92.

Patiño, Pedro. (2011). A specialized parallel corpus of English and Spanish Free Trade Agreements for the study of specialized collocations. *Synaps*, 26:85–89.

Patiño, Pedro. (2013). LSP in Colombia: advances and challenges. In Quiroz, G. and Patiño, Pedro, editors, *Toward a definition of specialized collocation*. Peter Lang, Bern.

Popel, Martin, Mareček, David, Green, Nathan, and Žabokrtsky, Zdeněk. (2011). Influence of parser choice on dependency-based MT. In *Proceedings of the 6th Workshop on Statistical Machine Translation*, pages 433–439, Edinburgh, Scotland. ACL.

Ramasamy, Loganathan and Žabokrtský, Zdeněk. (2011a).

Tamil dependency parsing: Results using rule based and corpus based approaches. In *Computational Linguistics and Intelligent Text Processing, 12th International Conference*, volume 6608 of *Lecture Notes in Computer Science*, pages 82–95, Berlin/Heidelberg. Springer-Verlag.

Ramasamy, Loganathan and Žabokrtský, Zdeněk. (2011b). Tamil Dependency Treebank (TamilTB) – Annotation manual. ÚFAL/CKL Technical Report TR-2011-42, Univerzita Karlova v Praze, Institute of Formal and Applied Linguistics.

Ramasamy, Loganathan and Žabokrtský, Zdeněk. (2012). Prague dependency style treebank for tamil. In *Proceedings of the 8th LREC*, Istanbul, Turkey. ELRA.

Ramasamy, Loganathan, Bojar, Ondřej, and Žabokrtský, Zdeněk. (2012a). Morphological processing for English-Tamil statistical machine translation. In *Proceedings of the Workshop on Machine Translation and Parsing in Indian Languages*, pages 113–122, Mumbai, India.

Ramasamy, Loganathan, Žabokrtský, Zdeněk, and Vajjala, Sowmya. (2012b). The study of effect of length in morphological segmentation of agglutinative languages. In *Proceedings of the 1st Workshop on Multilingual Modeling*, pages 18–24, Jeju, Republic of Korea. ACL.

Romeo, Lauren, Martínez Alonso, Héctor, and Bel, Núria. (2013). Class-based word sense induction for dot-type nominals. In *Proceedings of the 6th International Conference on Generative Approaches to the Lexicon*, Pisa, Italy.

Schumann, Anne-Kathrin. (2011a). A bilingual study of knowledge-rich context extraction in Russian and German. In *Proceedings of the 5th Language and Technology Conference*, pages 516–520. Fundacja Uniwersytetu im. A. Mickiewicza.

Schumann, Anne-Kathrin. (2011b). A case study of knowledge-rich context extraction in Russian. In *Proceedings of the 9th International Conference on Terminology and Artificial Intelligence*, pages 143–146, Paris. INALCO.

Schumann, Anne-Kathrin. (2011c). Extraction of knowledge-rich contexts in Russian – a study in the automotive domain. In *Proceedings of the 18th NoDaLiDa*, number 11 in NEALT Proceedings Series, pages 311–314.

Schumann, Anne-Kathrin. (2012a). Knowledge-rich context extraction and ranking with KnowPipe. In *Proceedings of the 8th LREC*, pages 3626–3630, Istanbul, Turkey. ELRA.

Schumann, Anne-Kathrin. (2012b). Towards the automated enrichment of multilingual terminology databases with knowledge-rich contexts. In *Proceedings of Dialogue 2012*, pages 559–567, Moscow, Russia. Российский государственный гуманитарный университет.

Schumann, Anne-Kathrin. (2012c). Towards the automated enrichment of multilingual terminology databases with knowledge-rich contexts – Experiments with Russian EuroTermBank data. In *Proceedings of the 2nd CHAT Workshop*, volume 72 of *Linköping Electronic Conference Proceedings*, pages 27–34.

Sulger, Sebastian, Butt, Miriam, King, Tracy Holloway, Meurer, Paul, Laczkó, Tibor, Rákosi, György, Dione, Cheikh Bamba, Dyvik, Helge, Rosén, Victoria, De Smedt, Koenraad, Patejuk, Agnieszka, Çetinoglu, Özlem, Arka, I Wayan, and Mistica, Meladel. (2013). ParGramBank: The ParGram parallel treebank. In *Proceedings of the 51st ACL*, volume 1, pages 550–560, Sofia, Bulgaria. ACL.

Susanto, Raymond Hendy, Larasati, Septina Dian, and Tyers, Francis M. (2012). Rule-based machine translation between Indonesian and Malaysian. In *Proceedings of the Workshop on South and Southeast Asian Natural Language Processing at COLING*, Mumbai, India.

Uria, Larraitz, Hulden, Mans, Etxeberria, Izaskun, and Alegria, Iñaki. (2011). Recursos y métodos de sustitución léxica en las variantes dialectales en euskera. In *Proceedings of the Workshop on Iberian Cross-Language NLP tasks*, pages 70–76. CEUR-WS.

Vajjala, Sowmya and Meurers, Detmar. (2012). On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–173, Montréal, Canada. ACL.

Vajjala, Sowmya and Meurers, Detmar. (2013). On the applicability of readability models to web texts. In *Proceedings of the 2nd Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 59–68, Sofia, Bulgaria. ACL.

Vajjala, Sowmya and Meurers, Detmar. (2014a). Exploring measures of "readability" for spoken language: Analyzing linguistic features of subtitles to identify age-specific tv programs. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations at EACL*, Gothenburg, Sweden. ACL.

Vajjala, Sowmya and Meurers, Detmar. (2014b). On assessing the reading level of individual sentences for text simplification. In *Proceedings of the 14th EACL Conference*, Gothenburg, Sweden. ACL.

Vajjala, Sowmya and Meurers, Detmar. (2014c). Readability assessment for text simplification: From analyzing documents to identifying sentential simplifications. *International Journal of Applied Linguistics, Special Issue on Current Research in Readability and Text Simplification*.

Wittenburg, Peter, Lenkiewicz, Przemyslaw, Auer, Erik, Lenkiewicz, Anna, Gebre, Binyam Gebrekidan, and Drude, Sebastian. (2012). AV processing in eHumanities – A paradigm shift. In *Proceedings of the Digital Humanities Conference*, pages 538–541, Hamburg, July.

Zampieri, Marcos and Gebre, Binyam Gebrekidan. (2012). Automatic identification of language varieties: The case of Portuguese. In *Proceedings of KONVENS2012*, pages 233–237, Vienna, Austria.

Zampieri, Marcos, Gebre, Binyam Gebrekidan, and Diwersy, Sascha. (2012). Classifying pluricentric languages: Extending the monolingual model. In *Proceedings of the 4th Swedish Language Technology Conference*, pages 79–80, Lund, Sweden.