LREC 2018
MIYAZAKI

ELEVENTH INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION

Held under the patronage of the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT)

MAY 7– 12, 2018

PHOENIX SEAGAIA CONFERENCE CENTRE
Miyazaki, JAPAN

CONFERENCE PROCEEDINGS

Editors: Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, Stelios Piperidis, Takenobu Tokunaga

Assistant Editors: Sara Goggi, Hélène Mazo

LREC 2018, ELEVENTH INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION

Title: LREC 2018 Conference Proceedings

Distributed by:
ELRA – European Language Resources Association
9, rue des Cordelières
75013 Paris
France

Tel.: +33 1 43 13 33 33
Fax: +33 1 43 13 33 30

www.elra.info and www.elda.org
Email: info@elda.org and lrec@elda.org

ISBN 979-10-95546-00-9
EAN 9791095546009

Introduction to LREC 2018 by Nicoletta Calzolari
Chair of the 11th edition of LREC
ELRA Honorary President

Welcome to the 11th edition of LREC in Miyazaki, first LREC in Asia!

LREC 20th Anniversary

It is the LREC 20th Anniversary and LREC has become one of the most successful conferences of the field. Data are pervasive in Natural Language Processing and Language Technology: we call our data Language Resources (LR). But when LREC was started by ELRA, in 1998 in Granada, from an idea of Antonio Zampolli and Joseph Mariani, it was really a new adventure and a challenge. There were well established big conferences but he thought that the new emerging field of Language Resources deserved its own dedicated forum. In the keynote talk I gave at LREC1998 I could say: "the infrastructural role of Language Resources as the necessary common platform on which new technologies and applications can be based is nowadays widely recognised." This could not have been said only few years before. I had the pleasure and the honour of being involved in LREC from the beginning, first as member of the Program Committee and since 2004 as Conference Chair.

LREC is probably the most influential ELRA achievement, and a service with the major impact on our community. Also through LREC, ELRA contributes to shape our field, making the Language Resource field a scientific field in its own right.

Why LREC in Asia this time? AFNLP (the Asian Federation of NLP) asked us if we could hold an LREC in Asia as the best instrument to promote Language Resources in Asia. We were glad to accept this challenge and here we are.

Some LREC2018 figures

As expected given the change in continent, we did not break any record this time, but the figures are not far from the previous. We received 1102 submissions for the main conference, 34 workshop proposals and 8 tutorial proposals.

A very large part of our community was involved in the reviewing effort, to be able to assign few papers per reviewer: 1263 colleagues accepted to act as reviewers (more than in 2016) out of 1796 invited (268 declined and 265 unfortunately not answering). Few reviewers did not complete the task (only 26 reviews missing, not so bad), but knowing that this always happens we recruited some pinch-reviewers able to act at the last moment: a good move to keep for the future.

The Program Committee has also been enlarged with 3 colleagues from Japan and one from USA. We had as usual a very hard job, examining about 3300 reviews, to understand – beyond the scores and in particular when they greatly differed – the relevance, the novelty, but also the appropriateness for an oral or poster presentation. I am sure we made mistakes, every reviewing effort is not immune from subjectivity, but as usual we discussed in a face to face meeting not only general policies, criteria and how to be consistent, but also borderline cases to arrive at agreed decisions. Overall we all believe we received in average good submissions. We have in the main program 718 papers: 188 Orals and 530 Posters.

We also have 29 Workshops and 5 Tutorials.

I am proud that around 1100 participants have already registered at the end of April, similar to last time. They come from 63 countries. The Japanese are the largest group and in general there is a larger participation from Asian countries, in particular China, as we obviously hoped.

These figures have a clear significance. The field of Language Resources and Evaluation is very alive and constantly flourishing.

LREC acceptance rate: a motivated choice for an inclusive conference

The LREC acceptance rate, 65% this year, is different from other major conferences but for us it is a motivated decision. This is one of the reasons why LREC succeeds to provide a comprehensive picture of the field and to show how it is evolving. For us it is important not only to hear about new methodologies but also to understand how various methods or resources are able to spread, for which purposes, usages, applications, and for which languages. Multilingualism – and equal treatment of all languages – is an essential feature of LREC, as it is the attempt of putting the text, speech and multimodal communities together as well as academics and industrials. LREC wants to be an "inclusive" conference.

Quality is not undermined by our acceptance rate: in 2017 Google Scholar Metrics h5-index, LREC ranks 4th in Computational Linguistics top conferences (5th considering ArXiv which is the first).

LREC2018 Novelties

Industry Track

Because of the interest in joining forces between academy and industry, this time we decided to experiment with a new Industry Track. We spoke about this at last LREC with Linne Ha from Google and we asked her if she wanted to organise it for LREC2018.

Special Speech Session

A special session on "Speech resources collection in real-world situations" was proposed to us by Kikuo Maekawa and Yuichi Ishimoto (National Institute for Japanese Language and Linguistics): we gladly accepted also to strengthen the participation of the speech community at LREC.

Oriental-COCOSDA Conference

Also O-COCOSDA is organised together with LREC. We spoke with Satoshi Nakamura, its chair, at last LREC and he kindly offered to organise it jointly with LREC. We are very pleased of this also because it is another opportunity to reach the Asian speech community.

ELRA Individual Members Assembly

ELRA has recently introduced "individual membership" in addition to institutional membership. This was decided to give a voice inside ELRA to the large LREC community and offer them its services. The first assembly of ELRA individual members is held on the first day of the conference.

The LREC Club

From the answers received, it seems that the LREC Club of those who attended all editions, the really faithful ones, is composed of 23 members. I want to thank them for their loyalty!

LREC2018 Trends

I quickly sketch here, as I always do, my perception – subjective and impressionistic – of LREC2018 trends and how certain topics fluctuate from an LREC to the other. The comparison with previous years shows the topics with steady progress, or even great leaps forward, the stable ones and those more affected by the fashion of the moment.

Trends in LREC2018 topics

Among the areas that continue to be trendy and are even increasing I can mention:

- Less-Resourced Languages
- Social Media analysis, appearing in 2012 and since then constantly growing
- Semantics in general and in particular Sentiment, Emotion and Subjectivity
- Information extraction, Knowledge discovery, Text mining are booming
- Lexicons (in its various forms)
- Discourse, Dialogue, Conversational systems and Interactivity
- Multimodality, also for Less-Resourced languages
- Tools, Systems, Applications for various purposes: Question Answering, Summarisation, etc.
- Evaluation methodologies
- Computer Aided Language Learning

Stable "usual" topics, some very well-represented, others in the medium/low range, are:

- Infrastructural issues, policies, strategies and Large projects: topics that receive special attention at LREC, differently from other major conferences
- Corpus creation, annotation, use, …
- Speech related topics, a little increasing but not as much as we would like
- Sign language (also a very successful workshop)
- Crowdsourcing
- Anaphora and Coreference
- Temporal and Spatial annotation

New trends for this LREC:

- Digital Humanities (new for LREC in 2016, now increased)
- Bibliometrics, Scientometrics, Infometrics
- Language Modelling

Decreasing topics with respect to the past, even if some still numerous:

- Grammar and syntax and also Treebanks that had a big increase in 2016
- Multilinguality and Machine Translation, very high in 2016
- Ontologies
- Standards and metadata are much less represented
- Linked data, a new topic in 2014, seems no longer so fashionable
- Web services and workflows also no longer so popular

The recognition given by the LR community to infrastructural issues, strategies and policies may be also due to the fact that we must often work in large groups, for many languages, we must build on each other work, connect various resources and tools, make available what already exists and use standardised formats. Infrastructures (on many dimensions) are really needed for our field to progress: to pay proper attention to these issues is another distinguishing feature of LREC.

LREC-related initiatives

Proceedings in Thomson Citation Index

Since 2010 the LREC Proceedings are included in CPCI (Thomson Reuters Conference Proceedings Citation Index): an important achievement, providing a better recognition to all LREC authors and useful in particular for young colleagues.

LRE Journal and LREC

After each LREC we ask to the authors of papers suggested by the 3 reviewers as appropriate for LRE if they want to submit an extended version to the *LRE journal*, coedited by Nancy Ide and myself. I am glad to report that also the journal has a large and increasing number of submissions, testifying the great interest for the field of LRs and Evaluation.

Citation of Language Resources

Also this year we encouraged citations of LRs in a special References section (introduced in 2016), providing recommendations on how to cite. I hope this becomes normal practice, to keep track of the relevance of LRs but also to provide due recognition to those working on LRs.

LRE Map and Share your Language Resources

As usual we encouraged descriptions of LRs in the LRE Map, an innovative instrument introduced at LREC2010 with the aim of monitoring the wealth of data and technologies developed and used in our field. And we ask, since 2014, to share the LRs with all the community.

In this LREC about 1000 LRs have been described in the Map. Just few hints at some data in the 2018 Map: WordNets, Wikipedia, Prague TreeBank are the most cited LRs; Corpora are by far the most frequent type (half of the LRs); and about 85% of the LRs are in some way available (not bad).

Replicability of research results

I believe that research is strongly affected also by infrastructural (meta-research) activities as those mentioned above. With these initiatives I hope we are able to promote in our field what is in use in more mature disciplines, i.e. ensure proper documentation and reproducibility of research results as a normal practice. ELRA and LREC are thus influential in strengthening the Language Resources and Evaluation scientific ecosystem and fostering sustainability.

Acknowledgments

As usual, it is my pleasure to express here my deepest gratitude to all those who made this LREC2018 possible and hopefully successful.

I first thank the Programme Committee members, not only for their dedication in the huge task of selecting the papers, the workshops and tutorials, but also for the constant involvement in the various aspects around LREC. I wish to thank each of them, the new ones:

Chris Cieri, Koiti Hasida, Hitoshi Isahara, Take Tokunaga, and obviously the “old” ones: Khalid Choukri, Thierry Declerck, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, Stelios Piperidis. A particular thanks goes to Jan Odijk, who has been so helpful in the preparation of the program.

I thank ELRA and the ELRA Board: LREC is a major service from ELRA to all the community!

A very special thanks goes to Sara Goggi and Hélène Mazo, chairs of the Editorial Committee, for the dedication and competence in managing the so many tasks they have in their hands, and the capability to tackle the many big and small problems of such a large conference (not an easy task). They are the two pillars of LREC, without whose commitment for many months LREC would not happen. So much of LREC organisation is on their shoulders, and this is visible to all participants.

I am especially grateful to Hitoshi Isahara and Kyoko Kanzaki, the Local Committee, for their great efforts in dealing with so many local matters and for their patience and true commitment, looking at so many details. We owe them a lot in organising a successful LREC.

My appreciation goes also to the distinguished members of the Advisory Board, chaired by Makoto Nagao, for their support and precious advices.

I am very grateful to the Local Liaison Committee, representing a number of Asian associations and organisations that have supported LREC, in particular with dissemination tasks.

I express my great gratitude to the Sponsorship Committee, and to all the Sponsors that have helped with financial support, believing in the importance of our conference.

I thank the Japanese Ministry of Education, Culture, Sports, Science and Technology for its precious support.

I am particularly grateful to the local authorities, the Governor of Miyazaki prefecture and the Mayor of Miyazaki city, very supportive of LREC since my first visit in 2016. We met them, and other local authorities, several times. We thank them also for their financial support.

In my many visits to Miyazaki I have been impressed by the great sense of hospitality of locals, and among them I wish to thank at least Manmatsu Hayashi and Rie Saita for their great help and kindness.

Also on behalf of the Program Committee, I praise our impressively large Scientific Committee. They did a wonderful job.

I thank the workshop and tutorial organisers, who complement LREC of so many interesting events.

I thank the organisers of the Industry Track, of the Special Speech session and of O-COCOSDA.

A big thanks goes to all the LREC authors, who provide the “substance” to LREC, and give us such a broad picture of the field.

This time I really want to thank also Softconf (Rich Gerber and Paolo Gai) and their constant efforts to make START a better tool for us. I greatly appreciated the new feature for plagiarism detection: I must say it was useful to detect some paper too similar to others …

I thank the European Commission for the interest in our conference, and hope that funding agencies will be impressed by the quality and quantity of initiatives in our sector that LREC displays, and by the fact that the field attracts the best groups of R&D from all continents. The success of LREC for us means the success of the field of Language Resources and Evaluation.

I finally thank the two institutions that always dedicate a great effort to LREC: ELDA in Paris and ILC-CNR in Pisa. Without their commitment LREC would not be possible. The last, but not least, thanks are thus, in addition to Hélène Mazo and Sara Goggi, to all the others who – with different roles – have helped and will help during the conference: Roberto Bartolini, Damien Bihel, Irene De Felice, Riccardo Del Gratta, Pawel Kamocki, Valérie Mapelli, Monica Monachini, Vincenzo Parrinelli, Vladimir Popescu, Valeria Quochi, Caroline Rannaud, Alexandre Sicard.

And lastly, my final words are for all the LREC2018 participants, the true protagonist of LREC. Now LREC is in your hands. I hope that you discover new paths, that you perceive the vitality and strength of the field, that you have fruitful conversations (conferences are useful also for this) and most of all that you profit of so many contacts to organise new exciting work and projects in the field of Language Resources and Evaluation … which you will show at the next LREC.

This Japanese LREC in Miyazaki has a sort of Mediterranean flavour, typical of LREC. I am sure you will appreciate the Japanese great hospitality and kindness. And I hope that Miyazaki will enjoy the invasion of LRECers!

With all the Programme Committee, I welcome you at LREC2018 and wish you a very fruitful Conference.

Enjoy LREC2018 in Miyazaki!

Message to LREC 2018 participants by Henk van den Heuvel ELRA President

Dear Colleagues and Friends,

Twenty years of LREC! It is my honour and pleasure to welcome you to this 11th edition of our successful Conference. Welcome to Miyazaki!

We are also very grateful with our guests representing the European Commission. Your presence here is deeply appreciated. Especially we welcome, Gael Kent, Director Data at the European Commission- DG CONNECT in Luxemburg. We are looking forward to your speech.

We are very honoured to have literally in our midst Prof. Makoto Nagao, Professor Emeritus of Kyoto University. We greatly admire your contributions to such various fields as Machine Translation, Natural Language Processing, Pattern Recognition, Image Processing and Library Science.

After 20 years LREC we have broken with the tradition to convene around the Mediterranean area, and look for another venue to meet and network. Honestly, we see this as an exceptional move motivated by our deep desire to intensify the ties with our Asian colleagues as we know them from the Asian Federation of Natural Language Processing (AFNLP), the Board of which is also closely involved in the organisation of this LREC through the Local Liaison Committee. We are very pleased to see so many of our Asian colleagues here in Miyazaki.

As President of ELRA it is my duty and pleasure to point out a couple of developments that are taking place in our Association. Already in 2012 one of my predecessors, Stelios Piperidis, referred to the dazzling speed of changes in which our community is finding itself. In his opening speech at LREC 2012 he also mentioned the upcoming of data-driven techniques and numerical and learning methods. In our days we see how algorithms and techniques developed in the area of Artificial Intelligence have come to play a paramount role in the area of Language and Speech technology. This technology puts special demands on the amount and preprocessing of Language Resources for training and testing purposes. Large amounts of data are collected from the web and continuously processed and used for application refinement. Now, in this rapidly changing field, ELRA has to find its way as one of the traditional sustainable key-players in language resources management and intermediary between stakeholders. It is evident that LRs remain essential also in our time, it is also evident that well-targeted annotated resources remain essential for supervised training approaches. Therefore, there remains an important role for ELRA as a sustainable LR broker offering relevant and high quality resources both to academia and commercial parties.

However, the changes that we see around us force us to continuously reflect on our *raison d'être* for our members in consideration of what their demands are for LRs and in terms of the services we offer around them. As a result of that, ELRA's Board has introduced important changes in its membership policy.

First of all, to stimulate continuity for our institutional members we have introduced a discount on membership fees upon membership continuation, starting with a discount of 15%

for the second year up to 30% for the third year and following. Second, we have equalized the fees for EU and non-EU members to the EU-members fee. Last but not least, as of January 2018 ELRA has introduced individual membership. An event such as LREC shows how vivid and productive the community around LRs is, and advocates for establishing a permanent link within this community, not only a biennial meeting point. For this reason ELRA has decided to open up its memberships for individuals, too, and to offer this membership with special services and benefits, of which the reduced registration fee is the one now most salient.

Employees of institutional ELRA members are also individual ELRA members if and when they want to use ELRA member services (including discount on LREC registration fees). They will not have to pay the individual membership fees as well since their organization covers for that.

In addition, one position in the ELRA Board will be reserved for a representative from the individual members, and this member is elected by the individual members only. This Board member has the same rights as the other ELRA Board members on all issues related to Board matters.

There will be a General Meeting for individual members at each LREC where they can convene with their representative and the Board to discuss ELRA matters concerning individual members. This meeting will be organized for the first time in this LREC 2018, namely this very afternoon at 18:00. The content of the meeting is an interesting mixture of relevant issues from the ELRA board, an inventory of wishes from individual members, and a self- introduction of Board applicants.

You are all invited to attend this first ELRA membership meeting, where we will tell more about the new membership policy, the special services for members and the election of the new Board member. We have sent out an invitation and an agenda for this.

Another observation that requires our persistent attention is that there are many players offering LRs both at the national and international level, and this landscape is becoming quite diffuse. This implies that we need to identify and re-identify times and again what our, ELRA's, position is compared to other LR brokers. It is ELRA's firm belief that this can best be done through cooperation. In this way we have set up a successful cooperation with for instance, LDC, by identifying the differences in membership policies, LR production and distribution strategies, and using each other's strengths in cooperation. In the same spirit ELRA has now set up a Collaboration Agreement with CLARIN ERIC. In this Collaboration Agreement we have clearly identified where our mutual and complementary strengths are and how we can bring these together to the benefit of both organisations. The objective of such agreements is not that one organization becomes part of the other but that both remain independent whilst joining forces. Indeed, here we see an important role for our association in facilitating synergies.

Another example of such a synergy has been established in our Special Interest Group for Under-resourced Languages, SIGUL. Created in April 2017, SIGUL is a joint Special Interest Group of the European Language Resources Association (ELRA) and of the International Speech Communication Association (ISCA). Through its establishment of the Special Interest Group on Under-resourced Languages, ELRA reasserts its active involvement in contributing to enhance the support for the languages with little or no technological support.

I would like to take the opportunity to thank all those who have worked so hard to make this conference a fantastic event: the LREC Programme Committee, chaired by Nicoletta Calzolari, the Scientific Committee, the Conference Editorial Committee headed by our LREC cornerstones Sara Goggi and Hélène Mazo, the International Advisory Committee chaired by Prof. Makoto Nagao, the group in Pisa, Khalid Choukri and the ELDA staff in Paris, the Local Committee headed by Prof. Hitoshi Isahara and Dr Kyoko Kanzaki. Each one of them in his/her own role has been taking care of the incredible amount of issues that emerge when undertaking the organisation of such a complex and demanding conference as LREC. Our particular thanks go to our sponsors and supporters.

We thank workshop and tutorial organizers, project consortia participating in the HLT Village; you have all exceeded yourselves once more to make this LREC such a great event.

Dear LREC Participants, in the end this is your conference. With your active participation in the oral sessions, your lively discussions with the presenters at the poster sessions, your visits to the HLT Village and Exhibition Boots and participation in the Industry Track I am confident you will make LREC 2018 yet another success.

Introductory message of Khalid Choukri,
ELRA Secretary General
ELDA Chief Executive Officer

ELRA and ELDA are very pleased to welcome you in Miyazaki to this 11th LREC to celebrate the 20th anniversary of LREC with all of you this week.

On behalf of the ELRA/ELDA team I would like to share with you some news on the activities we conducted since the last LREC in Portorož (Slovenia).

The Declaration of Granada

But first let me to share some feelings about this special LREC with you, as we are celebrating the 20th anniversary of this major forum established in 1998 in Granada (Spain), organized for its 11th edition, here in Japan.

Soon after the establishment of ELRA in 1995, its Board realised that, at that time, the language resources and the evaluation of language technologies were given very little attention at the main events. Today, we are glad that such message is spread widely and is endorsed by the major conferences in which special sessions are expressly devoted to Language Resources and Evaluation!!

Remember the first LREC, remember Granada, not only the Alhambra! With over 400 participants instead of the expected 100 attendees, we realized the importance of such forum for the community. This was confirmed over the years by a steady attendance of 1200 participants to the last editions of LREC.

I would like to take this opportunity to go back to the spirit of Granada, paying a tribute to those who were behind it, Professors Antonio Zampolli and Angel Martin Municio. I would like to bring up one of the major outcomes of that first event: "the declaration of Granada". Its recommendations are still relevant and topical, more urgent than ever to implement.

The declaration of Granada[1] comprised 10 articles. I am highlighting and commenting here some of the crucial ones that we can continue to endorse today:

- "**At this moment, language resources are one indispensable key to unlock the potential of the global information Society**"

We are still facing this issue 20 years later and if we agree that the Information Society has made tremendous progress with the emergence of social networks which have strengthened links within and between communities, social or commercial activities cross borders are still hindered by language barriers. In 2015, surveys mentioned that 24 languages are used in LinkedIn user interfaces, 48 on Twitter, 91 on Google Translate (as pairs for its translation of content and now about 103), over 150 on Facebook, just over 300 in Wikipedia. These

[1] Granada Declaration: http://www.elra.info/media/filer_public/2013/09/06/v3n3.pdf)

numbers may seem impressive, but remember that this is **out of 7097 living languages or 3,909 with writing systems.** And most of these languages are used in interfaces with automatic processing of content used in Search and/or MT only. Language Resources are essential assets. Back in Granada, we stated that "They constitute an essential infrastructure". Such infrastructure is missing for a huge number of languages. The LRE Map service provided by ELRA, inventorying the LRs reported in major conferences, continue to expose the existing gaps.

- **"All sectors of society, and all languages, have an interest in seeing these resources developed, for a variety of purposes, economic, social, industrial and cultural."**

ELRA continues to promote the concept of Basic Language Resource Kit, a Kit that would help process every language for (at least) the basic NLP functions. We stressed the importance of this approach to policy makers, emphasised the need to support small communities, and mentioned the lack of interest from private sector for non-lucrative/non-strategic languages. We also insisted that such *"core language resources should remain in the public domain*" to ensure a wide use by both research and development stakeholders. Reviewing the current situation at major data centers and repositories, we can barely count more than 100 different languages, often with scarce resources (many speech resources for the major languages, very few treebanks, very few aligned corpora, mostly aligned with English, etc.)

- **"For each language, there is a need for strategy to co-ordinate existing resources and create new ones."**

ELRA, along with LDC, their partner in the USA, did their best to offer distribution/sharing channels for Language Resources produced within publicly funded projects and some offered by private bodies. However the identified resources represent less than 15% of what exists. Coordination of the distribution but also documentation and production, have proved to be challenging. We still feel it is crucial to coordinate building roadmaps for every language and enhance the involvement of local public and private bodies. It is also essential to continue international cooperation to disseminate the know-how acquired for a given language. We are glad that a conference like LREC contributes to sharing such expertise and value the implication of governmental (regional and national) and international bodies.

We introduced the International Standard Language Resource Number (now part of the activities of the International Standardisation Organisation, ISO TC37/SC4) to assign a unique identifier with each identified Language Resource to improve the way we reference it (this is also part of the LREC submission process that distinguishes Bibliographical data from LR data). The idea is not only to provide an ID, unique and persistent, wherever the LR is stored, even for those LRs on local servers outside the Internet. This is an uphill struggle but we are convinced that it is an important step in our work to improve the identification of existing resources, the assessment of LR impact factor as well as the citation mechanism.

- **"When resources have been created, there is a continuing requirement for support and maintenance."**

This is a key part of our mission and we tried to convince data producers and funders to account for the necessary maintenance of and support for Language Resources. We introduced the validation process and the "bug" reporting mechanism, as part of ELRA procedures, to encourage sharing experiences on the use of LRs and their enhancement over time. We still face funding scenarios that provide subsidies for data production and not for other issues like IPR clearance, documentation, sharing, maintaining, etc. In Granada, we anticipated that resources would undergo some repurposing with the new uses that emerge and we insisted on the need to envisage a wide range of applications on the basis of the same resources. The community seems to be sensitive to this, but some legislators are debating the adoption of more legal constraints. We need to join forces to convince funders and decision makers about the importance of more openness and long term policies. The introduction of the Data Management Plan (DMP) by ELRA, and soon the DMP Wizard, will help each data manager to adopt up-to-date standards and best practices for data management.

- **"Understanding of the role, usefulness and optimum means of preparation for language resources is a research theme in itself."**

Over the last decades, and especially within the last 3-4 years, we have seen an impressive breakthrough in the HLT field. The new data-intensive machine learning and the computing capabilities, are proving the crucial usefulness of LRs. Making LRs widely available is the core mission of a few organisations. ELRA is very happy to be among these organizations and is making the necessary investments to acquire more expertise to cost-effectively produce and share LRs. The setup of an internal legal team is helping to shed light on a large number of legal issues that impede the use/re-use of LRs. Working on standards is also an important aspect to help facilitate the interoperability and sharing of data. One of our mottos was that "Common evaluation requires common standards". We still feel that common tasks in the "challenges" and evaluation campaigns are essential instruments to assess progress, share knowledge, and improve cooperation. It is a pity that many "Evaluation campaigns" are happening with very little coordination which makes them hard to find for new comers.

Granada was 20 years ago and we see that some visionary recommendations are still needed today. A multilateral, concrete, and lasting cooperation remains on top of our action.

ELRA activities since 2016

Now allow me to get back to ELRA activities carried out over the last couple of years.

We continue our actions on data sharing, through the identification, negotiation, and distribution agreements with right holders when necessary. We continue to produce resources for projects as well as for partners. Our policy remains consistent: whenever the data is offered to the community, after the shortest possible embargo period, the costs for partners are set to production costs. This position remains fundamental to our policy. We continue to invest in research and development of tools to improve and automate our production procedures. Most of our tools are shared as open source packages.

We continue also to work on our quality control methodologies so as to supply validated resources with validation procedures that guarantee the adequacy of the produced datasets with respect to the initial specifications and the state of the art.

To ensure an efficient distribution of Language Resources, ELRA has migrated its catalogue of resources to a new platform, based on e-Commerce features, redesigned with a new interface and an improved navigation. This foreshadows further developments that will incorporate e-licensing, e-payment and e-delivery of resources.

ELRA continues to support the set-up of LR repositories for data deposit by third parties. Based on its involvement in the jointly-developed META-SHARE platform, we continue the promotion of such efforts to ensure that the major data holders adhere to some common practices. A new repository was set up as part of an EU service contract to store data for MT provided by the public sector. Such initiative is now spreading across Europe, and a coordination action is establishing local repositories (known as Local Relay Stations). If we succeed to set up such stations for each country in order to collect all language datasets produced by translations services and secure these for MT training and tuning, one can anticipate good progress for these languages and domains. The repositories can accommodate any Language Resource modality.

If the establishment of such a local repository is of interest to your organization and your network, let us discuss how to work on it together.

As part of this process, we continue to work on all issues related to sustainability and preservation of data for the generations to come.

An updated ELRA Data Management Plan is made available and reviews all necessary aspects for an optimal management of resources with an easy-to-use checklist. We are working to automate the customisation of such DMP for each project. Our members will benefit from this automatic DMP Wizard, accompanied with the support of our experts, free of charge. We hope that such approach will improve sustainability and preservation of Language Resources but also make them easy to identify.

ELRA continues to be involved in the new trends in HLTs. It continues to support the new trends in MT. Many of our projects (some of which are funded under a European Program known as Connecting Europe Facility (CEF) focus on data production, including via requests for donations from translation services, but also crawling of adequate data to which we have access and re-use rights. Many resources come from organizations that belong to the Public Sector. A directive (called Public Sector Information directive, PSI) entered into application in the European Union, similar rules exist in many other countries, stating that publicly produced data should be made publicly available. This makes some of the resources needed by our community (e.g. textual corpora) available for new domains and new genres. Some geographical areas offer a multilingual environment (EU, India? South Africa, etc.), and hence more resources should be available for MT development.

Unfortunately there are still important legal restrictions on the re-use of data, even for research purposes. We continue to vilify the current legal framework, in particular in Europe, e.g. the European Union is working on a new directive on copyright in the Digital Single Market. The initial proposal for this act contained a mandatory exception for text and data mining carried out by research institutions. However, the current debates within the European decision makers seem to suggest that the exception will fall short of meeting the objective of the exception. The beneficiaries of the new exception may be limited to public research institutions, and – more importantly – ‘lawful access’ will be a prerequisite for data mining, which will probably result in wider implementation of digital protection measures by right holders. It is unlikely to get the exception for research that we claim since years now as a fair use doctrine for research purposes (that remains the privilege of a few countries).

The current legal framework has a strong impact on the capacity of the community to produce IPR cleared and sharable data. ELRA heavily invested in legal training and has been, for many years now, one of the few organizations that works both with in-house legal experts and a network of external practitioners/lawyers.

Another critical novelty in Europe is the new legal framework governing the processing of personal data. It goes beyond the users expectations, for more ethical behaviour on the management of their data. This may hinder the new developments of resources and technologies (e.g. Crowdsourcing activities). The new regulation (General Data Protection Regulation (GDPR)) will impose more restrictions on managing several aspects of data e.g. data protection by design and by default, privacy impact assessment, pseudonymisation and anonymization, before the data can be shared (this will of course impact also production, repackaging, repurposing of data).

To share information on these matters, a dedicated workshop on legal and ethical issues continues to be organized within LREC and will be held this week as well.

Of course, ELRA does not focus on EU issues and EU languages only (we distribute resources for more than 70 languages). In 2017, ELRA entered into an important agreement with the International Speech Communication Association (ISCA[2]) to join forces in the promotion of activities related to the Less-Resourced Languages (LRL). ELRA and ISCA agreed to merge their groups and set up a join Special Interest Group for Under-resourced Languages (SIGUL[3]). Co-chaired by a representative of ELRA and a representative of ISCA, SIGUL will continue to organize events for the LRL and encourage cooperation actions to support these languages.

As you may know, United Nation General Assembly proclaimed 2019 as the International Year of indigenous Languages. UNESCO is leading the corresponding events. ELRA proposed to organize an important international event related to HLT and Indigenous languages. We hope to draw attention to the importance of HLT and LRs for the preservation and development of local cultures and put under spotlights the role our community could play for these languages.

We continue to develop the LRE Map application. LRE Map was established to reference all LRs described by authors when submitting papers to conferences and journals. Started with LREC, it is used by other events but not as widely as we hope. In addition to identifying over 7000 instances of LRs, it helps identify existing gaps for languages lacking such modalities and ensure a minimal cooperation when planning new productions. If you are involved in the organisation of a conference, let us see how we can work together.

ELRA is also taking part in several standardisation activities. It is naturally involved in ISO/TC37/SC 4 on Language Resource management but also on ISO/IEC JTC1/SC35 about user interfaces and accessibility. ELRA brings its knowledge of the HLT field to ensure that all ICT services and products are accessible to all, in particular to users with specific needs. Some of the HLT applications are offering valuable services when converting speech into text, text into speech, sub-titling/captioning audio-visual streams, providing audio descriptions, translations (e.g. subtitles), easy-reading features (both in mono- and multilingual contexts). Such services are valuable to everyone and not only hearing or visually impaired users. Translation from text or speech to Sign languages is a big challenge that many partners are working on and ELRA will support them.

As a conclusion to my message, I would like to reiterate my statement uttered at almost all LRECs since 1998. Please remember that we can help you share your data for all types of use. We can work out a contractual framework that suits your expectations, including adopting very permissive licences and a free-of-charge policy. We can guarantee the availability as well

[2] https://www.isca-speech.org/iscaweb/index.php/about-isca

[3] http://www.elra.info/en/sig/sigul/

as the sustainability of your resources. During the conference, an ELRA booth is available where we will be happy to interact with you on such topics.

About 10 years ago, we identified about 20 resources, some were on the web, others well known to the community. We keep monitoring their availability. Believe it or not, about 30% disappeared and these are not necessarily the ones that were obsolete and useless. Some right holders also disappeared and the "orphan" resources with them.

Acknowledgments

Finally, I would like to express my deep thanks to our partners and supporters, who throughout the years make LREC so successful. I would like to thank our Sponsors: Google, Amazon, Arcadia, EML (European Media Laboratory GmbH), GSK Language Resources, Riken-AIP, Yahoo Research Japan, and the publisher Hituzi Syobo our media sponsor: MultiLingual Computing, Inc.

I also would like to thank the HLT Village participants, we hope that such gathering offers the projects an opportunity to foster their dissemination and hopefully to discuss exploitation plans with the participants.

I would like to thank the Local Advisory Committee. Its composition of the most distinguished personalities of Japan denotes the importance of language and language technologies for the country.

I would like to thank the LREC Local Committee, chaired by Prof. Hitoshi Isahara and the LREC Local Organizing Committee, for providing support to the organization of this LREC Edition in Japan.

Finally I would like to warmly thank the joint team of the two institutions that devoted so much effort over months and often behind curtains to make this one week memorable: ILC-CNR in Pisa and my own team, ELDA, in Paris. These are the two LREC coordinators and pillars: Sara Goggi and Hélène Mazo, and the team: Roberto Bartolini, Damien Bihel, Irene De Felice, Valérie Mapelli, Monica Monachini, Vincenzo Parrinelli, Vladimir Popescu, Caroline Rannaud, and Alexandre Sicard.

Now LREC 2018 is yours: we hope that each of you will achieve valuable results and accomplishments. We, ELRA and ILC-CNR staff, are at your disposal to help you get the best out of it.

Once again, welcome to Miyazaki and Japan, welcome to LREC 2018

Message to LREC 2018 participants by of Hitoshi Isahara and Kyoko Kanzaki, Chairs of the Local Committee

Welcome to Japan! Welcome to Miyazaki!!

On behalf of the whole local team, we would like to extend our warmest welcome to all of you participating in this 11th edition of LREC in Miyazaki, Japan!

You may know personally some Japanese researchers in the fields which are involved in LREC, and also may know that Japan is a country with four distinctive seasons, blessed with beautiful nature, world heritage sites, rich culture and respected traditions. However, we are sure that most of you are not familiar with Miyazaki.

Miyazaki has beautiful coastline facing the Pacific Ocean, and has many shrines involved with the myth of the birth of Japan, such as Miyazaki Shrine sacred to Emperor Jinmu, supposedly the first Emperor of Japan. Most stories in the mythology associated with the creation of Japan and the origin of the imperial line took place in Miyazaki Prefecture on the island of Kyushu.

You can visit scenic places and historical places in Miyazaki in this occasion, or you will be able to visit here again with your friends and family.

During your stay here, we would like you to experience Omatsuri, festival in Japan. You can enjoy Shrine Maiden Dance, Kagura performance (a sacred music and dancing performance dedicated to the Shinto gods) and local cuisine during reception. Lunch will be served in stall style.

On this occasion, we would like to thank all the supporters in Japan from the bottom of **our** hearts.

We first appreciate the patronage of the Ministry of Education, Culture, Sports, Science and Technology (MEXT) towards LREC2018.

We thank Miyazaki Prefecture and Miyazaki city for their continuous support from the beginning of LREC's venue selection process. We offer the Governor, Mr. Shunji Kono, and the Mayor, Mr. Tadashi Tojiki our heartfelt thanks for their great kindness and efforts. Thanks to their generous understandings, preparation of LREC went smoothly.

We also thank Miyazaki Convention and Visitors Bureau for its financial and human support. We are grateful to Mr. Mitsunori Mera, Mr. Toshiaki Tomitaka and Mr. Manmatsu Hayashi.

We are sure that all participants satisfy high quality service of the conference venue, Phoenix Seagaia Resort. We appreciate very hard work by Mr. Hirofumi Matsunaga, Mr. Manmatsu Hayashi, Ms. Rie Saita and Mr. Satoru Kamibayashi to meet the requirements which LREC indicated.

One of the highlights of LREC2018 is the Reception at Miyazaki Shrine which is sacred to the first Emperor of Japan. We are grateful to its Chief Priest, Mr. Hidekiyo Sugita, for accepting us.

Participants can enjoy local liquor at the Reception and Gala Dinner. Let's thank to Miyazaki Sake Brewers Association.

We would like to thank Ms. Rika Kubota for organizing interpreter volunteers during LREC.

We would like to thank the Japan National Tourism Organization (JNTO) which supported us during LREC's venue selection process, including site visit to choose the best place for LREC in Japan.

Lastly, we would like to thank Mr. Manmatsu Hayashi for his great effort to make LREC success. The word “impossible” couldn't be found in his dictionary.

We are ready to welcoming you with *omotenashi*, our traditional spirit of hospitality.

Enjoy LREC2018 in Miyazaki!

LREC 2018 Committees

Conference Programme Committee

Nicoletta Calzolari	**ILC/CNR, Pisa, Italy (Conference chair)**
Khalid Choukri	ELRA, Paris, France
Christopher Cieri	LDC, Philadelphia, USA
Thierry Declerck	DFKI GmbH, Saarbrücken, Germany
Koiti Hasida	The University of Tokyo, Tokyo, Japan
Hitoshi Isahara	Toyohashi University of Technology, Toyohashi, Japan
Bente Maegaard	CST, University of Copenhagen, Denmark
Joseph Mariani	LIMSI-CNRS, Orsay, France
Jan Odijk	UIL-OTS, Utrecht, The Netherlands
Asuncion Moreno	Universitat Politècnica de Catalunya, Barcelona, Spain
Stelios Piperidis	ILSP, Athens, Greece
Takenobu Tokunaga	Tokyo Institute of Technology, Tokyo, Japan

Advisory Board

Shyam S. Agrawal	KIIT, Gurgaon (India)
Hiroya Fujisaki	University of Tokyo (Japan)
Eva Hajičová	UFAL, Charles University, Prague (Czech Republic)
Yuming Li	Beijing Language and Culture University (PRC)
Mark Liberman	Linguistic Data Consortium, Philadelphia (USA)
Makoto Nagao (Chair)	**Professor Emeritus University of Kyoto (Japan)**
Jun'ichi Tsujii	Artificial Intelligence Research Center, Tokyo (Japan)

Local Liaison Committee

Key-Sun Choi	KAIST (Korea)
Chu-Ren Huang	The Hong Kong Polytechnic University (Hong Kong SAR - PRC)
Toru Ishida	Department of Social Informatics, Kyoto University (Japan)
Haizhou Li	National University of Singapore (Singapore)
Satoshi Nakamura	Nara Institute of Science and Technology (Japan)
Byong-Rae Ryu	Chungnam National University (Korea)
Virach Sornlertlamvanich	Sirindhorn International Institute of Technology, Thammasat University (Thailand)
Le Sun	Chinese Academy of Sciences, Beijing (PRC)
Kam-Fai Wong	The Chinese University of Hong Kong (Hong Kong SAR - PRC)
Chengqing Zong	Chinese Academy of Sciences, Beijing (PRC)

Scientific Committee

The Programme Committee is very grateful to Scientific Committee members who reviewed the submissions and contributed to designing the conference programme. The list of the members of Scientific Committee is published on the LREC 2018 web site.

- **Local Committee**

Hitoshi Isahara (Chair)	Toyohashi University of Technology, Toyohashi, Japan
Kyoko Kanzaki	Toyohashi University of Technology, Toyohashi, Japan

- **Conference Editorial Committee**

Sara Goggi	ILC/CNR, Pisa, Italy
Hélène Mazo	ELDA/ELRA, Paris, France

- **Organising Committee**

Roberto Bartolini	ILC/CNR, Pisa, Italy
Damien Bihel	ELDA/ELRA, Paris, France
Irene De Felice	University of Pisa, Italy
Riccardo Del Gratta	ILC/CNR, Pisa, Italy
Sara Goggi	**ILC/CNR, Pisa, Italy (Co-chair)**
Valérie Mapelli	ELDA/ELRA, Paris, France
Hélène Mazo	**ELDA/ELRA, Paris, France (Co-chair)**
Monica Monachini	ILC/CNR, Pisa, Italy
Vincenzo Parrinelli	ILC/CNR, Pisa, Italy
Vladimir Popescu	ELDA/ELRA, Paris, France
Valeria Quochi	ILC/CNR, Pisa, Italy
Caroline Rannaud	ELDA/ELRA, Paris, France
Alexandre Sicard	ELDA/ELRA, Paris, France

- **Sponsorship Committee**

Nicoletta Calzolari	ILC/CNR, Pisa, Italy
Khalid Choukri	ELRA, Paris, France
Tatjana Gornostaja	Tilde, Riga, Latvia
Hitoshi Isahara	Toyohashi University of Technology, Toyohashi, Japan
Kyoko Kanzaki	Toyohashi University of Technology, Toyohashi, Japan
Jimmy Kunzmann	EML GmbH, Heidelberg, Germany
Joseph Mariani	LIMSI-CNRS & IMMI, Orsay, France
Satoshi Sekine	New York University, New York City, USA

Acknowledgements

The European Language Resources Association, ELRA, and the LREC Committees acknowledge with gratitude the support and sponsoring of the following institutions.

Sponsors and supporters

- Google *(Platinum)*
- AIP, Center for Advanced Intelligent Project *(Bronze*)
- Amazon AWS *(Bronze*)
- Arcadia Computing Innovation *(Bronze*)
- EML European Media Laboratory *(Bronze*)
- Yahoo Research Japan *(Bronze)*
- GSK Language Resources Association *(Supporter)*
- Hituzi Syobo *(Publisher*)
- Multilingual (*Media Sponsor*)

Supporting Institutions

- Evaluations and Language resources Distribution Agency (ELDA), Paris (France)
- Istituto di Linguistica Computazionale (ILC) of the Italian National Research Council (CNR), Pisa (Italy)
- Toyohashi University of Technology, Japan
- Miyazaki Convention and Visitors Bureau

Designing a Collaborative Process to Create Bilingual Dictionaries of Indonesian Ethnic Languages

[1]Arbi Haza Nasution, [2]Yohei Murakami, [3]Toru Ishida
[1,3]Department of Social Informatics, Kyoto University, [2]Unit of Design, Kyoto University
Kyoto, Japan
[1]arbi@ai.soc.i.kyoto-u.ac.jp, [2]yohei@i.kyoto-u.ac.jp, [3]ishida@i.kyoto-u.ac.jp

Abstract

The constraint-based approach has been proven useful for inducing bilingual dictionary for closely-related low-resource languages. When we want to create multiple bilingual dictionaries linking several languages, we need to consider manual creation by a native speaker if there are no available machine-readable dictionaries are available as input. To overcome the difficulty in planning the creation of bilingual dictionaries, the consideration of various methods and costs, plan optimization is essential. Utilizing both constraint-based approach and plan optimizer, we design a collaborative process for creating 10 bilingual dictionaries from every combination of 5 languages, i.e., Indonesian, Malay, Minangkabau, Javanese, and Sundanese. We further design an online collaborative dictionary generation to bridge spatial gap between native speakers. We define a heuristic plan that only utilizes manual investment by the native speaker to evaluate our optimal plan with total cost as an evaluation metric. The optimal plan outperformed the heuristic plan with a 63.3% cost reduction.

Keywords: Bilingual Dictionary Creation, Low-resource Languages, Closely-related Languages

1. Introduction

Nowadays, machine-readable bilingual dictionaries are being utilized in actual services (Ishida, 2011) to support intercultural collaboration (Ishida, 2016; Nasution et al., 2017b), but low-resource languages lack such sources. Obviously bilingual lexicon extraction is highly problematic for low-resource languages due to the paucity or outright omission of parallel and comparable corpora. We introduced the promising approach of treating pivot-based bilingual dictionary induction for low-resource languages as an optimization problem (Nasution et al., 2016; Nasution et al., 2017c) where bilingual dictionaries are the only language resource required. Despite the high potential of our approach in enriching low-resource languages, it faces numerous issues when trying to create plans to implement multiple bilingual dictionaries for a set of low-resource languages like Indonesian ethnic languages. When actually implementing our constraint-based bilingual dictionary induction approach, we need to consider the inclusion of more traditional methods like manually creating the bilingual dictionaries by native speaker. In spite of the high cost, this will be unavoidable if no machine-readable dictionaries are available. Given the various methods and costs that may need to be considered, we recently introduced a plan optimizer to find the feasible optimal plan of creating multiple bilingual dictionaries with the least total cost (Nasution et al., 2017a). In this project, to create bilingual dictionary D_{A-B} between ethnic language L_A and ethnic language L_B, there is also a difficulty in finding a bilingual native speaker of two ethnic languages. To overcome this limitation, we can firstly create triple T_{A-ID-B} using the common language, Indonesian as pivot language L_{ID} where S_{ID-A}, a native bilingual speaker of Indonesian language L_{ID} - ethnic language L_A and S_{ID-B}, a native bilingual speaker of Indonesian language L_{ID} - ethnic language L_B collaborate by explaining the senses with Indonesian language. Then, the bilingual dictionary D_{A-B} can be induced from the triple T_{A-ID-B}. The native speakers need a tool that can bridge the spatial gap and help them collaborate. To actually implement our pivot-based bilingual dictionary induction following the optimal plan to create multiple Indonesian ethnic languages bilingual dictionaries, we address the following research goals:

- *Designing a Collaborative Process for Creating Bilingual Dictionaries of Indonesian Ethnic Languages*: Implementing plan optimization for creating bilingual dictionaries of low-resource languages and implementing a generalized constraint approach to bilingual dictionary induction for low-resource language families in creating 10 bilingual dictionaries with 2,000 translation pairs from every combination of 5 languages, i.e., Indonesian, Malay, Minangkabau, Javanese, and Sundanese.
- *Designing an Online Collaborative Dictionary Generation*: Bridging spatial gap between native speakers especially when doing a collaborative creation or evaluation of bilingual dictionary.

The rest of this paper is organized as follows: In Section 2 and Section 3, we will briefly discuss our constraint-based bilingual dictionary induction and plan optimizer, respectively. Section 4 details our collaborative process design. Finally, Section 5 concludes this paper.

2. Constraint-Based Bilingual Dictionary Induction

The traditional pivot-based approach is very suitable for low-resource languages (Tanaka and Umemura, 1994). Unfortunately, for some low-resource languages, it is often difficult to find machine-readable inverse dictionaries and corpora to identify and eliminate the erroneous translation

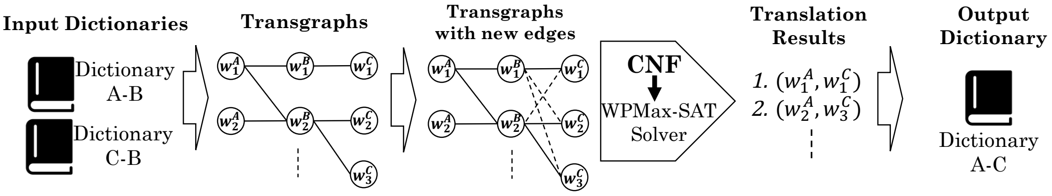


Figure 1: One-to-one constraint approach to pivot-based bilingual dictionary induction.

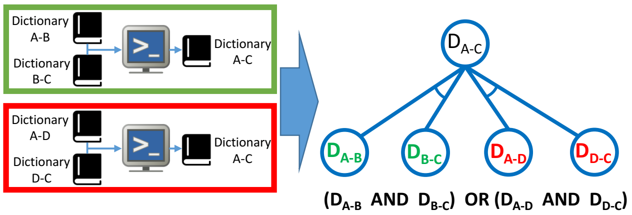


Figure 2: Modeling Bilingual Dictionary Induction Dependency.

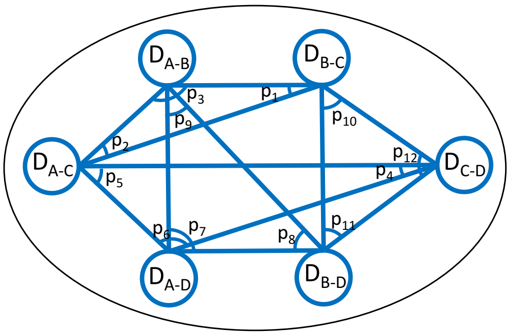

D_{A-B}
D_{B-C}
D_{A-C}
D_{C-D}
D_{A-D}
D_{B-D}
p_1 p_2 p_3 p_4 p_5 p_6 p_7 p_8 p_9 p_{10} p_{11} p_{12}

Figure 3: AND/OR Graph as an MDP State.

pair candidates. To overcome this limitation, our team (Wushouer et al., 2015) proposed to treat pivot-based bilingual lexicon induction as an optimization problem. The assumption was that lexicons of closely-related languages offer instances of one-to-one mapping and share a significant number of cognates (words with similar spelling/form and meaning originating from the same root language). The proposal uses a graph whose vertices represent words and edges indicate shared meanings; following (Soderland et al., 2009) it was called a transgraph. The proposal proceeds as follows: (1) use two bilingual dictionaries as input, (2) represent them as transgraphs where w_1^A and w_2^A are non-pivot words in language L_A, w_1^B and w_2^B are pivot words in language L_B, and w_1^C, w_2^C and w_3^C are non-pivot words in language L_C, (3) add some new edges represented by dashed edges based on the one-to-one assumption, (4) formalize the problem into conjunctive normal form (CNF) and use the Weighted Partial MaxSAT (WPMaxSAT) solver (Ansótegui et al., 2009) to return the optimized translation results, and (5) output the induced bilingual dictionary as the result. These steps are shown in Figure 1. However, the assumption of one-to-one mapping is too strong to induce the many-to-many translation pairs needed to offset resource paucity because few such pairs can be found. Therefore, we generalized the constraint-based bilingual dictionary induction framework by extending constraints and translation pair candidates from the one-to-one approach to attain more voluminous bilingual dictionary results with many-to-many translation pairs extracted from connected existing and new edges (Nasution et al., 2016). We further enhance our generalized method by setting two steps to obtaining translation pair results. First, we identify one-to-one cognates by incorporating more constraints and heuristics to improve the quality of the translation result. We then identify the cognates' synonyms to obtain many-to-many translation pairs. In each step, we can obtain more cognate and cognate synonym pair candidates by iterating the n-cycle symmetry assumption until all possible translation pair candidates have been reached (Nasution et al., 2017c).

3. Plan Optimizer

Our constraint-based bilingual dictionary induction approach has the potential to enrich low-resource languages with the only input being machine readable bilingual dictionaries. Unfortunately, the scarcity of such dictionaries for low-resource languages makes it difficult to plan which bilingual dictionary should be invested first or which bilingual dictionary should be induced right from the start in order to obtain all possible combination of bilingual dictionaries from the language set with the minimum total cost to be paid. We model the bilingual dictionary dependency with AND/OR graphs as shown in Figure 2, and employ the Markov Decision Process (MDP) for plan optimization where a state is defined by AND/OR graphs as shown in Figure 3. The exponential complexity of formulating the bilingual dictionary creation planning into a graph theory problem indicates a greater complexity of obtaining the optimal planning with the least total cost by only following the heuristic. Nevertheless, our algorithm greatly reduced the complexity, so that the MDP planning can find the feasible optimal plan with less total cost compared to heuristic planning (e.g., only use manual investment by native speaker). Our MDP model can calculate the cumulative cost while predicting and considering the probability of the pivot action to yield a satisfying output bilingual dictionary as utility for every state to better predict the most feasible optimal plan with the least total cost. Our formalization with MDP allow user to predict the feasible optimal plan with the least total cost before implementing the constraint-based bilingual dictionary induction framework in a big scale.

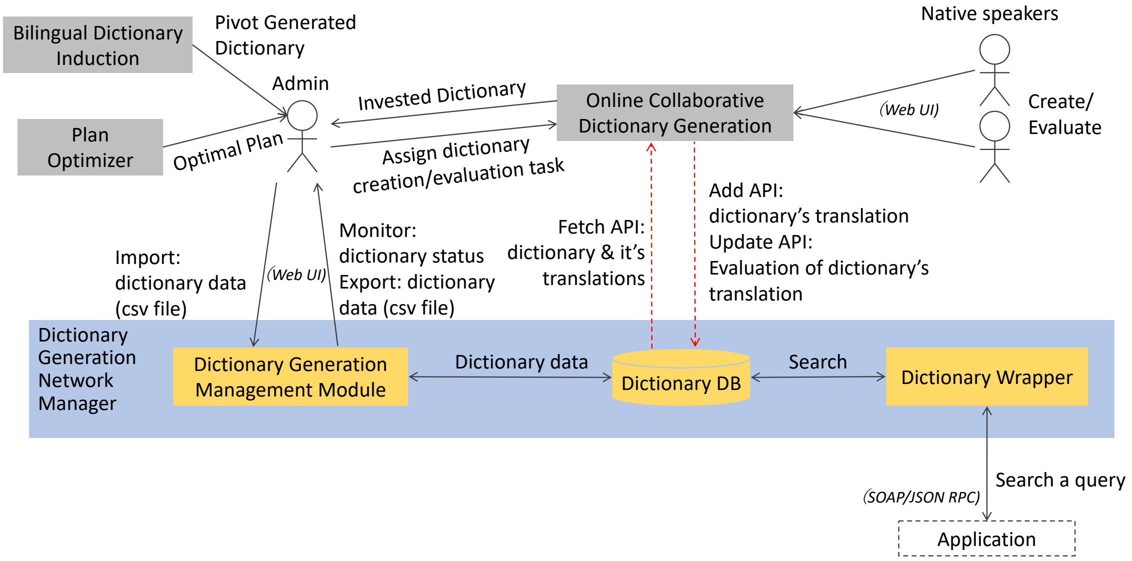

Bilingual Dictionary Induction
Pivot Generated Dictionary
Admin
Plan Optimizer
Optimal Plan
Invested Dictionary
Assign dictionary creation/evaluation task
Online Collaborative Dictionary Generation
Native speakers
(Web UI)
Create/ Evaluate
Import: dictionary data (csv file)
(Web UI)
Monitor: dictionary status
Export: dictionary data (csv file)
Fetch API: dictionary & it's translations
Add API: dictionary's translation
Update API: Evaluation of dictionary's translation
Dictionary Generation Network Manager
Dictionary Generation Management Module
Dictionary data
Dictionary DB
Search
Dictionary Wrapper
Search a query
(SOAP/JSON RPC)
Application

Figure 4: System Integration Overview.

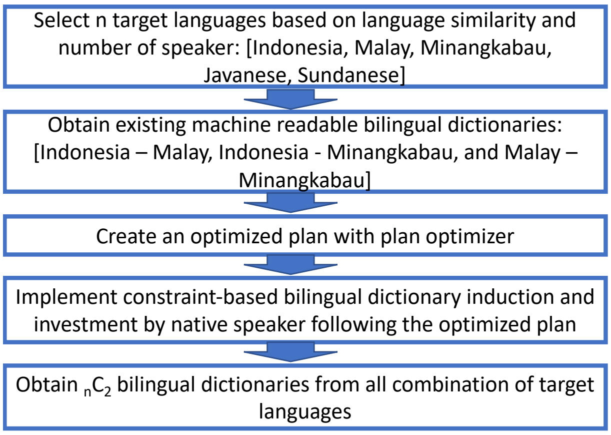

Select n target languages based on language similarity and number of speaker: [Indonesia, Malay, Minangkabau, Javanese, Sundanese]
Obtain existing machine readable bilingual dictionaries: [Indonesia – Malay, Indonesia - Minangkabau, and Malay – Minangkabau]
Create an optimized plan with plan optimizer
Implement constraint-based bilingual dictionary induction and investment by native speaker following the optimized plan
Obtain nC2 bilingual dictionaries from all combination of target languages

Figure 5: Overview of Bilingual Dictionaries Generation Process.

4. Designing a Collaborative Process

4.1. Overview

We integrate our Constraint-based Bilingual Dictionary Induction and Plan Optimizer with an Online Collaborative Dictionary Generation as a tool to bridge the spacial gap between native speakers and a Dictionary Generation Network Manager to manage the final dictionary so that it is accessible via API in the Language Grid (Ishida, 2011) as shown in Figure 4. The overview of bilingual dictionaries generation process is shown in Figure 5 while the detailed process is explained in Algorithm 1.

4.2. Selecting Target Languages

To select target languages in this paper, we use an Automatic Similarity Judgment Program (ASJP) (Holman et al., 2011) following our previous work (Nasution et al., 2017d). Indonesia has 707 low-resource ethnic languages (Lewis et al., 2015) that require our attention. There are two factors we consider in selecting the target languages: language similarity and number of speakers. In order to ensure that the induced bilingual dictionaries will be useful for many users, we listed the top 10 Indonesian ethnic languages ranked by the number of speakers. Since our constraint-based approach works better on closely related languages, we further generated the language similarity matrix by utilizing ASJP as shown in Table 1. Based on number of speaker, we select Javanese and Sundanese. To find and coordinate native speakers of those languages, we collaborate with Telkom University. Based on relatedness with Indonesian, we select Malay and Minangkabau. To find and coordinate native speakers of those language, we collaborate with Islamic University of Riau. Hence, we target 5 languages, i.e., Indonesian, Malay, Minangkabau, Javanese, and Sundanese. We want to enrich/create the following dictionaries: Indonesia-Malay, Indonesia-Minangkabau, Indonesia-Javanese, Indonesia-Sundanese, Malay-Minangkabau, Malay-Javanese, Malay-Sundanese, Minangkabau-Javanese, Minangkabau-Sundanese, and Javanese-Sundanese with 2,000 translation pairs each.

4.3. Modeling Task for Native Speaker

When actually implementing our constraint-based bilingual dictionary induction approach, we need native speakers for manual creation of bilingual dictionaries or evaluation of the output dictionaries. There are a lot of prior researches on modeling workflow management (Georgakopoulos et al., 1995; Hollingsworth and Hampshire, 1995; Kappel et al., 2000; Huang et al., 2000; Alexopoulos et al., 2011; Kulkarni et al., 2012). We define several rules of which native speaker can create/evaluate which dictionary.

A bilingual dictionary between ethnic language L_A and ethnic language L_B, D_{A-B} can be induced from a triple T_{A-ID-B}, while a triple T_{A-ID-B} can be induced from a bilingual dictionary D_{ID-A} and a bilingual dictionary D_{ID-B}. A bilingual dictionary between Indonesian language L_{ID} and ethnic language L_A, D_{ID-A} can be man-

Algorithm 1: Bilingual Dictionaries Generation

```
Input: targetLanguageInfo, existingDictionaries
/* In this project, targetLanguages: [Indonesia,Malay,Minangkabau,Javanese,Sundanese] */
/* targetLanguageInfo includes language similarities and expectedDictionarySize=2,000 */
/* existingDictionaries=[$D_{Indonesia-Malay}$, $D_{Indonesia-Minangkabau}$, $D_{Malay-Minangkabau}$] */
Output: dictionaryList /* all combination of bilingual dictionaries from the targetLanguages */
for each $D_{A-B}$ in existingDictionaries do
    dictionaryList.add($D_{A-B}$);
end
optimizedPlan ← planOptimizer.create(targetLanguageInfo, dictionaryList);
for each action to create bilingual dictionary $D_{A-B}$ in optimizedPlan do
    if final state is reached then
        return dictionaryList
    end
    else
        if action type = investment then
            /* CT1($L_{ID}$,$L_A$): Creation and Evaluation of Indonesia-Ethnic Bilingual Dict */
            if $L_A$ or $L_B$ is Indonesian language $L_{ID}$ then
                create and evaluate bilingual dictionary $D_{A-B}$ by a native bilingual speaker $S_{A-B}$;
                dictionaryList.add($D_{A-B}$);
            end
            /* CT2($L_A$,$L_B$): Creation and Evaluation of Ethnic-Ethnic Bilingual Dict */
            else
                if native bilingual speaker $S_{A-B}$ is available then
                    create and evaluate bilingual dictionary $D_{A-B}$ by a native bilingual speaker $S_{A-B}$;
                    dictionaryList.add($D_{A-B}$);
                end
                else
                    create and evaluate triple $T_{A-ID-B}$ by two native bilingual speakers $S_{ID-A}$ and $S_{ID-B}$;
                    induce $D_{A-B}$ from $T_{A-ID-B}$; dictionaryList.add($D_{A-B}$);
                end
            end
        end
        else if action type = pivot then
            use constraint-based approach to obtain triple $T_{A-P-B}$ ;
            /* T4($L_A$,$L_P$,$L_B$) */
            if native bilingual speaker $S_{A-B}$ is available then
                evaluate triple $T_{A-P-B}$ by a native bilingual speaker $S_{A-B}$;
                induce $D_{A-B}$ from $T_{A-P-B}$; dictionaryList.add($D_{A-B}$);
            end
            else
                evaluate triple $T_{A-P-B}$ by two native bilingual speakers $S_{ID-A}$ and $S_{ID-B}$;
                induce $D_{A-B}$ from $T_{A-P-B}$; dictionaryList.add($D_{A-B}$);
            end
        end
    end
end
```

ually created or evaluated by a native bilingual speaker S_{ID-A}. A bilingual dictionary D_{A-B} can be manually created or evaluated by a native bilingual speaker S_{ID-A} and a native bilingual speaker S_{ID-B} collaboratively or by a native bilingual speaker S_{A-B} alone.

There are some bilingual dictionaries between Indonesian and Indonesian ethnic languages exist in a printed format. We may be able to digitalized the printed Indonesian - ethnic language bilingual dictionaries to a machine readable format. Nevertheless, when we connect the digitalized bilingual dictionary D_{ID-A} and a bilingual dictionary D_{ID-B} via Indonesian language L_{ID} as a pivot, and further induced D_{A-B} with our constraint-based approach, we expect that there will be many unreachable translation pair candidates since some Indonesian words in one bilingual dictionary may not exist in the other bilingual dictionary. In order to maximize the use of our pivot-based approach, we prepare a list of 2,000 most commonly used Indonesian words to be translated to ethnic language L_A to create a bilingual dictionary D_{ID-A} by a native bilingual speaker S_{ID-A} as shown in Figure 6. Due to budget limitation, we only allow the native speaker to translate an Indonesian word to up to five words of ethnic language L_A. To ensure the quality of the manually created bilingual dictionary D_{ID-A}, another native bilingual speaker S_{ID-A} will evaluate the translation pairs as shown

Table 1: Similarity Matrix of Top 10 Indonesian Ethnic Languages Ranked by Number of Speakers

Language	**Indonesian**	Malang	Yogyakarta	**Javanese**	**Sundanese**	**Malay**	Palembang Malay	Madurese	**Minangkabau**
Malang	23.46%								
Yogyakarta	27.29%	87.36%							
Javanese	**24.09%**	47.50%	52.18%						
Sundanese	**39.43%**	18.55%	22.43%	**21.82%**					
Malay	**85.10%**	20.53%	24.35%	**21.36%**	**41.12%**				
Palembang Malay	68.24%	33.97%	37.97%	31.85%	38.90%	73.23%			
Madurese	34.45%	17.63%	14.15%	15.18%	19.86%	34.16%	34.32%		
Minangkabau	**61.59%**	26.59%	29.63%	**25.01%**	**30.81%**	**61.66%**	63.60%	34.32%	
Buginese	31.21%	12.76%	16.85%	18.33%	24.80%	32.04%	31.00%	17.94%	32.00%

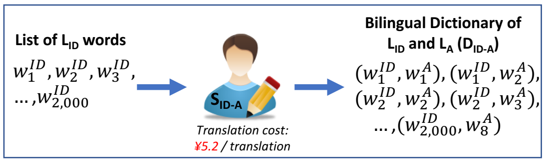

List of L_{ID} words
$w_1^{ID}, w_2^{ID}, w_3^{ID}, \ldots, w_{2,000}^{ID}$
S_{ID-A}
Translation cost:
¥5.2 / translation
Bilingual Dictionary of L_{ID} and L_A (D_{ID-A})
$(w_1^{ID}, w_1^A), (w_1^{ID}, w_2^A), (w_2^{ID}, w_2^A), (w_2^{ID}, w_3^A), \ldots, (w_{2,000}^{ID}, w_8^A)$

Figure 6: $T1(L_{ID}, L_A)$: Creation of Bilingual Dictionary D_{ID-A}.

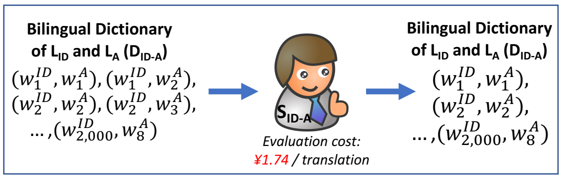

Bilingual Dictionary of L_{ID} and L_A (D_{ID-A})
$(w_1^{ID}, w_1^A), (w_1^{ID}, w_2^A), (w_2^{ID}, w_2^A), (w_2^{ID}, w_3^A), \ldots, (w_{2,000}^{ID}, w_8^A)$
S_{ID-A}
Evaluation cost:
¥1.74 / translation
Bilingual Dictionary of L_{ID} and L_A (D_{ID-A})
$(w_1^{ID}, w_1^A), (w_2^{ID}, w_2^A), \ldots, (w_{2,000}^{ID}, w_8^A)$

Figure 7: $T2(L_{ID}, L_A)$: Evaluation of Bilingual Dictionary D_{ID-A}.

in Figure 7. To overcome the limitation in finding native bilingual speakers of two ethnic languages for creation and evaluation of bilingual dictionary D_{A-B}, two native bilingual speakers S_{ID-A} and S_{ID-B} can collaborate as shown in Figure 8 and Figure 9 respectively. Finally, there are two composite tasks, which are $CT1(L_{ID}, L_A)$, a manual creation followed by evaluation of bilingual dictionary D_{ID-A} as shown in Figure 10a and $CT2(L_A, L_{ID}, L_B)$, a manual creation followed by evaluation of bilingual dictionary D_{A-B} as shown in Figure 10b.

4.4. Online Collaborative Dictionary Generation

The online collaborative dictionary generation has 6 modules: individual creation of Indonesia-Ethnic bilingual dictionary, individual evaluation of Indonesia-ethnic bilingual dictionary, individual creation of ethnic-ethnic bilingual dictionary, individual evaluation of ethnic-ethnic bilingual dictionary, collaborative creation of ethnic-ethnic bilingual dictionary, and collaborative evaluation of ethnic-ethnic bilingual dictionary. Each native speakers get his/her own user account. They can login to the system, read the user manual, update their profile, check their assigned task, and do their assigned task. For the individual task, the native speakers can do the task anywhere before the deadline as shown in Figure 11. However, for the collaborative task, a pair of native speakers need to login to the system at the same time in order to collaborate. The live chat is used to ease communication and discussion during the collaborative creation / evaluation session as shown in Figure 12.

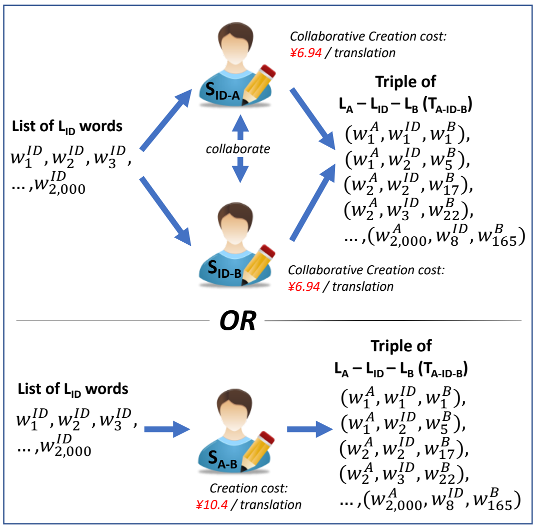

List of L_{ID} words
$w_1^{ID}, w_2^{ID}, w_3^{ID}, \ldots, w_{2,000}^{ID}$
S_{ID-A}
Collaborative Creation cost:
¥6.94 / translation
collaborate
S_{ID-B}
Collaborative Creation cost:
¥6.94 / translation
Triple of $L_A - L_{ID} - L_B$ (T_{A-ID-B})
$(w_1^A, w_1^{ID}, w_1^B), (w_1^A, w_1^{ID}, w_5^B), (w_2^A, w_2^{ID}, w_{17}^B), (w_2^A, w_3^{ID}, w_{22}^B), \ldots, (w_{2,000}^A, w_8^{ID}, w_{165}^B)$
OR
List of L_{ID} words
$w_1^{ID}, w_2^{ID}, w_3^{ID}, \ldots, w_{2,000}^{ID}$
S_{A-B}
Creation cost:
¥10.4 / translation
Triple of $L_A - L_{ID} - L_B$ (T_{A-ID-B})
$(w_1^A, w_1^{ID}, w_1^B), (w_1^A, w_2^{ID}, w_5^B), (w_2^A, w_2^{ID}, w_{17}^B), (w_2^A, w_3^{ID}, w_{22}^B), \ldots, (w_{2,000}^A, w_8^{ID}, w_{165}^B)$

Figure 8: $T3(L_A, L_{ID}, L_B)$: (Individual/Collaborative) Creation of Triple T_{A-ID-B} to induce Bilingual Dictionary D_{A-B}.

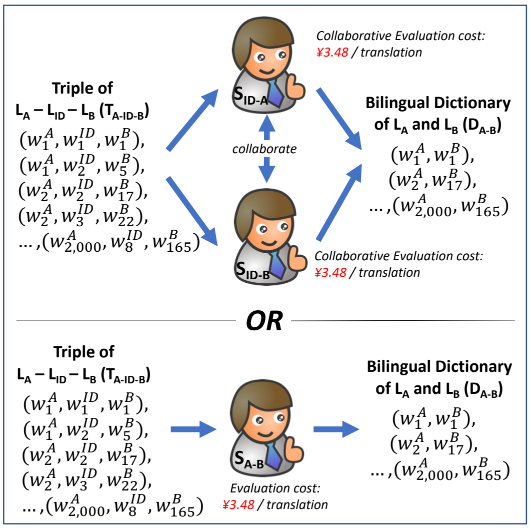

Triple of $L_A - L_{ID} - L_B$ (T_{A-ID-B})
$(w_1^A, w_1^{ID}, w_1^B), (w_1^A, w_2^{ID}, w_5^B), (w_2^A, w_2^{ID}, w_{17}^B), (w_2^A, w_3^{ID}, w_{22}^B), \ldots, (w_{2,000}^A, w_8^{ID}, w_{165}^B)$
S_{ID-A}
Collaborative Evaluation cost:
¥3.48 / translation
collaborate
S_{ID-B}
Collaborative Evaluation cost:
¥3.48 / translation
Bilingual Dictionary of L_A and L_B (D_{A-B})
$(w_1^A, w_1^B), (w_2^A, w_{17}^B), \ldots, (w_{2,000}^A, w_{165}^B)$
OR
Triple of $L_A - L_{ID} - L_B$ (T_{A-ID-B})
$(w_1^A, w_1^{ID}, w_1^B), (w_1^A, w_2^{ID}, w_5^B), (w_2^A, w_2^{ID}, w_{17}^B), (w_2^A, w_3^{ID}, w_{22}^B), \ldots, (w_{2,000}^A, w_8^{ID}, w_{165}^B)$
S_{A-B}
Evaluation cost:
¥3.48 / translation
Bilingual Dictionary of L_A and L_B (D_{A-B})
$(w_1^A, w_1^B), (w_2^A, w_{17}^B), \ldots, (w_{2,000}^A, w_{165}^B)$

Figure 9: $T4(L_A, L_{ID}, L_B)$: (Individual/Collaborative) Evaluation of Triple T_{A-ID-B} to induce Bilingual Dictionary D_{A-B}.

Table 2: Estimated Cost of Actions following MDP Optimal Plan

Task following MDP Plan	#Translation[1]	MDP Transition Probability[2]	Estimated Precision[2]	Unit Cost (JPY)	Total Cost (JPY)
T1(Indonesian, Malay)	1,480[3]			5.20	7,696.00
T2(Indonesian, Malay)	1,480			1.74	2,575.00
T1(Indonesian, Javanese)	2,000			5.20	10,400.00
T2(Indonesian, Javanese)	2,000			1.74	3,480.00
T1(Indonesian, Sundanese)	2,000			5.20	10,400.00
T2(Indonesian, Sundanese)	2,000			1.74	3,480.00
P(Malay, Indonesia, Minangkabau)	1,645[3]	0.983	0.4113	0.00	0.00
T4(Malay, Indonesian, Minangkabau)	754			6.96	5,248.00
P(Javanese, Indonesian, Sundanese)	1,027	0.972	0.2567	0.00	0.00
T4(Javanese, Indonesian, Sundanese)	1,027			6.96	7,147.00
T3(Javanese, Sundanese)	973			13.88	13,507.00
T4(Javanese, Sundanese)	973			6.96	6,773.00
P(Malay, Indonesia, Javanese)	1,094	0.943	0.2481	0.00	0.00
T4(Malay, Indonesia, Javanese)	1,094			6.96	7,615.00
T3(Malay, Javanese)	906			13.88	12,575.00
T4(Malay, Javanese)	906			6.96	6,305.00
P(Minangkabau, Indonesian, Sundanese)	1,157	0.949	0.289	0.00	0.00
T4(Minangkabau, Indonesian, Sundanese)	1,157			6.96	8,049.00
T3(Minangkabau, Sundanese)	844			13.88	11,708.00
T4(Minangkabau, Sundanese)	844			6.96	5,871.00
P(Malay, Indonesian, Sundanese)	1,356	0.826	0.3045	0.00	0.00
T4(Malay, Indonesian, Sundanese)	1,356			6.96	9,434.00
T3(Malay, Sundanese)	645			13.88	8,946.00
T4(Malay, Sundanese)	645			6.96	4,486.00
P(Minangkabau, Malay, Javanese)	1,148	0.929	0.2608	0.00	0.00
T4(Minangkabau, Malay, Javanese)	1,148			6.96	7,993.00
T3(Minangkabau, Javanese)	852			13.88	11,820.00
T4(Minangkabau, Javanese)	852			6.96	5,927.00
TOTAL					**171,435.00**

[1] *A number of translations is calculated from the number of translation pair candidates from the constraint-based approach × estimated precision with a high polysemy rate.*

[2] *Estimated from beta distribution based on language similarity and high polysemy pivot rate following our unpublished ACM TALLIP article entitled ”Plan Optimization to Bilingual Dictionary Induction for Low-Resource Language Families”.*

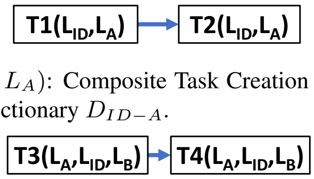
[3] *Excluding translation pairs from existing bilingual dictionaries: Indonesian-Malay (520 translation pairs) and Malay-Minangkabau (1,246 translation pairs).*

table

Table 3: Estimated Cost of Actions following Heuristic Plan

Task following Heuristic Plan	#Translation[1]	Unit Cost (JPY)	Total Cost (JPY)
T1(Indonesian, Javanese)	2,000	5.20	10,400.00
T2(Indonesian, Javanese)	2,000	1.74	3,480.00
T1(Indonesian, Sundanese)	2,000	5.20	10,400.00
T2(Indonesian, Sundanese)	2,000	1.74	3,480.00
T1(Indonesian, Malay)	1,480[1]	5.20	7,696.00
T2(Indonesian, Malay)	1,480	1.74	2,575.20
T3(Javanese, Sundanese)	2,000	13.88	27,760.00
T4(Javanese, Sundanese)	2,000	6.96	13,920.00
T3(Malay, Minangkabau)	754[1]	13.88	10,465.52
T4(Malay, Minangkabau)	2,000	6.96	13,920.00
T3(Malay, Javanese)	2,000	13.88	27,760.00
T4(Malay, Javanese)	2,000	6.96	13,920.00
T3(Minangkabau, Sundanese)	2,000	13.88	27,760.00
T4(Minangkabau, Sundanese)	2,000	6.96	13,920.00
T3(Malay, Sundanese)	2,000	13.88	27,760.00
T4(Malay, Sundanese)	2,000	6.96	13,920.00
T3(Minangkabau, Javanese)	2,000	13.88	27,760.00
T4(Minangkabau, Javanese)	2,000	6.96	13,920.00
TOTAL			**270,816.72**

[1] *Excluding translation pairs from existing bilingual dictionaries: Indonesian-Malay (520 translation pairs) and Malay-Minangkabau (1,246 translation pairs).*



(a) $CT1(L_{ID}, L_A)$: Composite Task Creation and Evaluation of Bilingual Dictionary D_{ID-A}.

(b) $CT2(L_A, L_{ID}, L_B)$: Composite Task Creation and Evaluation of Bilingual Dictionary D_{A-B}.

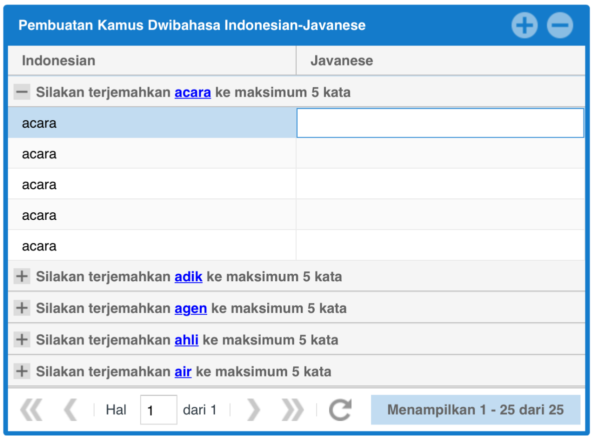
Figure 10: Composite Tasks.

4.5. Cost Estimation

We estimate the cost of each native speaker tasks as follows:

- $T1(L_{ID}, L_A)$: From an estimated duration of 30 seconds per translation and a daily wage of JPY5,000/8 hours, the estimated total translation per day is $1 \times 2 \times 60 \times 8 = 960$ and the estimated cost is JPY5.2 per correct translation.

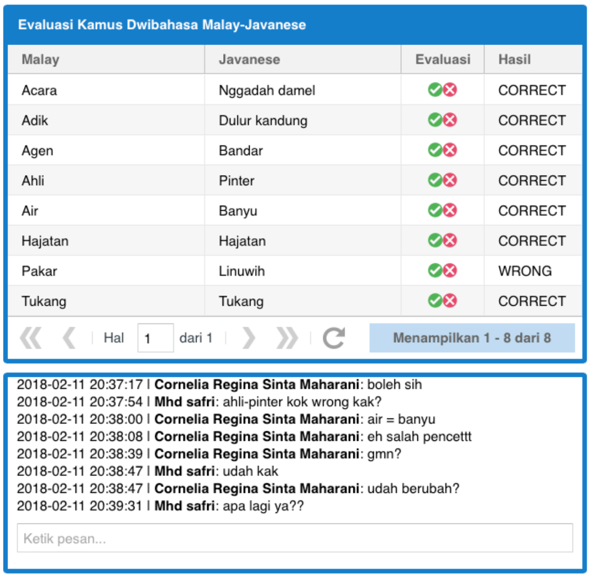
- $T2(L_{ID}, L_A)$: From an estimated duration of 10 seconds per translation and a daily wage of JPY5,000/8 hours, the estimated total translation per day is $1 \times 6 \times 60 \times 8 = 2,880$ and the estimated cost is JPY1.74 per



Pembuatan Kamus Dwibahasa Indonesian-Javanese

Indonesian	Javanese
− Silakan terjemahkan acara ke maksimum 5 kata	
acara	
acara	
acara	
acara	
acara	
+ Silakan terjemahkan adik ke maksimum 5 kata	
+ Silakan terjemahkan agen ke maksimum 5 kata	
+ Silakan terjemahkan ahli ke maksimum 5 kata	
+ Silakan terjemahkan air ke maksimum 5 kata	

Figure 11: Individual Creation of Indonesia-Ethnic Bilingual Dictionary.



Evaluasi Kamus Dwibahasa Malay-Javanese

Malay	Javanese	Evaluasi	Hasil
Acara	Nggadah damel	✓✗	CORRECT
Adik	Dulur kandung	✓✗	CORRECT
Agen	Bandar	✓✗	CORRECT
Ahli	Pinter	✓✗	CORRECT
Air	Banyu	✓✗	CORRECT
Hajatan	Hajatan	✓✗	CORRECT
Pakar	Linuwih	✓✗	WRONG
Tukang	Tukang	✓✗	CORRECT

2018-02-11 20:37:17 | **Cornelia Regina Sinta Maharani**: boleh sih
2018-02-11 20:37:54 | **Mhd safri**: ahli-pinter kok wrong kak?
2018-02-11 20:38:00 | **Cornelia Regina Sinta Maharani**: air = banyu
2018-02-11 20:38:08 | **Cornelia Regina Sinta Maharani**: eh salah pencettt
2018-02-11 20:38:39 | **Cornelia Regina Sinta Maharani**: gmn?
2018-02-11 20:38:47 | **Mhd safri**: udah kak
2018-02-11 20:38:47 | **Cornelia Regina Sinta Maharani**: udah berubah?
2018-02-11 20:39:31 | **Mhd safri**: apa lagi ya??

Figure 12: Collaborative Evaluation of Ethnic-Ethnic Bilingual Dictionary.

correct translation.

- $T3(L_A, L_{ID}, L_B)$: Following the cost of $T1(L_{ID}, L_A)$ and $T2(L_{ID}, L_A)$, for the individual task, from an estimated duration of 60 seconds per translation, the estimated cost is JPY5.2×2 = JPY10.4 per translation. For the collaborative task, from an estimated duration of 30 seconds to translate an Indonesian word to each ethnic language in parallel, and an extra 10 seconds for discussing the sense sharing between the two ethnic language translations, the estimated total cost is (JPY5.2 + JPY11.74)×2 workers = JPY13.88 per correct translation pair.

- $T4(L_A, L_{ID}, L_B)$: Following the cost of $T1(L_{ID}, L_A)$ and $T2(L_{ID}, L_A)$, for the individual task, from an estimated duration of 20 seconds per translation, the estimated cost is JPY1.74×2 = JPY3.48 per translation. For the collaborative task, from an estimated duration of 20 seconds to evaluate by discussing the sense sharing between the two ethnic language translations, the estimated total cost is (JPY1.74 + JPY1.74)×2 workers = JPY6.96 per correct translation pair.

- $CT1(L_{ID}, L_A)$: Following the cost of $T1(L_{ID}, L_A)$ and $T2(L_{ID}, L_A)$, the estimated cost is JPY5.2 + JPY1.74 = JPY6.94 per translation.

- $CT2(L_A, L_B)$: Following the cost of $T3(L_A, L_{ID}, L_B)$ and $T4(L_A, L_{ID}, L_B)$ and the combination of workers based on availability of native bilingual speakers ($S_{A-B} + S_{A-B}$, $S_{A-B} + S_{ID-A}\&S_{ID-B}$, $S_{ID-A}\&S_{ID-B} + S_{A-B}$, $S_{ID-A}\&S_{ID-B} + S_{ID-A}\&S_{ID-B}$), the variations of estimated total cost are (JPY10.4 + JPY3.48 = JPY13.88, JPY10.4 + JPY6.96 = JPY17.36, JPY13.88 + JPY3.48 = JPY17.36, JPY13.88 + JPY6.96 = JPY20.84) respectively.

We estimate the cost of actions following the optimized plan utilizing both constraint-based approach and manual investment by native speakers as shown in Table 2 and further compare them with cost of actions following the heuristic plan utilizing only manual investment by native speakers as shown in Table 3.

5. Conclusion

We design a collaborative process for creating 10 bilingual dictionaries with 2,000 translation pairs from every combination of 5 languages, i.e., Indonesian, Malay, Minangkabau, Javanese, and Sundanese. We implement our plan optimizer and our generalized constraint approach to bilingual dictionary induction in creating input dictionaries or evaluating the resulting bilingual dictionaries. We define a heuristic plan that only utilize manual investment by native speaker to evaluate our optimal plan with total cost as an evaluation metric. By following the optimal plan, we can reduce 63.3% cost of following the heuristic plan. We further design an online dictionary generation tool to bridge spatial gap between native speakers. We will analyze the native speakers' behavior and chat log for future improvement of the system.

6. Acknowledgment

This research was partially supported by a Grant-in-Aid for Scientific Research (A) (17H00759, 2017-2020) and a Grant-in-Aid for Young Scientists (A) (17H04706, 2017-2020) from Japan Society for the Promotion of Science (JSPS). The first author was supported by Indonesia Endownment Fund for Education (LPDP). We thank our collaborators, Department of Informatics Engineering, Universitas Islam Riau and Department of Computational Science, Telkom University for their effort in gathering the student workers.

7. Bibliographical References

Alexopoulos, K., Makris, S., Xanthakis, V., and Chryssolouris, G. (2011). A web-services oriented workflow management system for integrated digital production engineering. *CIRP Journal of Manufacturing Science and Technology*, 4(3):290 – 295. Production Networks Sustainability.

Ansótegui, C., Bonet, M. L., and Levy, J. (2009). Solving (weighted) partial maxsat through satisfiability testing. In *Theory and Applications of Satisfiability Testing-SAT 2009*, pages 427–440. Springer.

Georgakopoulos, D., Hornick, M., and Sheth, A. (1995). An overview of workflow management: From process modeling to workflow automation infrastructure. *Distributed and Parallel Databases*, 3(2):119–153, Apr.

Hollingsworth, D. and Hampshire, U. (1995). Workflow management coalition: The workflow reference model. *Document Number TC00-1003*, 19.

Holman, E. W., Brown, C. H., Wichmann, S., Müller, A., Velupillai, V., Hammarström, H., Sauppe, S., Jung, H., Bakker, D., Brown, P., et al. (2011). Automated dating of the world's language families based on lexical similarity. *Current Anthropology*, 52(6):841–875.

Huang, G., Huang, J., and Mak, K. (2000). Agent-based workflow management in collaborative product development on the internet. *Computer-Aided Design*, 32(2):133 – 144.

Toru Ishida, editor. (2011). *The Language Grid: Service-Oriented Collective Intelligence for Language Resource Interoperability*. Springer Publishing Company, Incorporated.

Ishida, T. (2016). Intercultural collaboration and support systems: A brief history. In *International Conference on Principles and Practice of Multi-Agent Systems (PRIMA 2016)*, pages 3–19. Springer.

Kappel, G., Rausch-Schott, S., and Retschitzegger, W. (2000). A framework for workflow management systems based on objects, rules and roles. *ACM Comput. Surv.*, 32(1es), March.

Kulkarni, A., Can, M., and Hartmann, B. (2012). Collaboratively crowdsourcing workflows with turkomatic. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW '12, pages 1003–1012, New York, NY, USA. ACM.

M. Paul Lewis, et al., editors. (2015). *Ethnologue: Languages of the World*. SIL International, Dallas, Texas, 18th edition.

Nasution, A. H., Murakami, Y., and Ishida, T. (2016). Constraint-based bilingual lexicon induction for closely related languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3291–3298, Paris, France, May.

Nasution, A. H., Murakami, Y., and Ishida, T. (2017a). Plan optimization for creating bilingual dictionaries of low-resource languages. In *2017 International Conference on Culture and Computing (Culture and Computing)*, pages 35–41, Sept.

Nasution, A. H., Syafitri, N., Setiawan, P. R., and Suryani, D. (2017b). Pivot-based hybrid machine translation to support multilingual communication. In *2017 International Conference on Culture and Computing (Culture and Computing)*, pages 147–148, Sept.

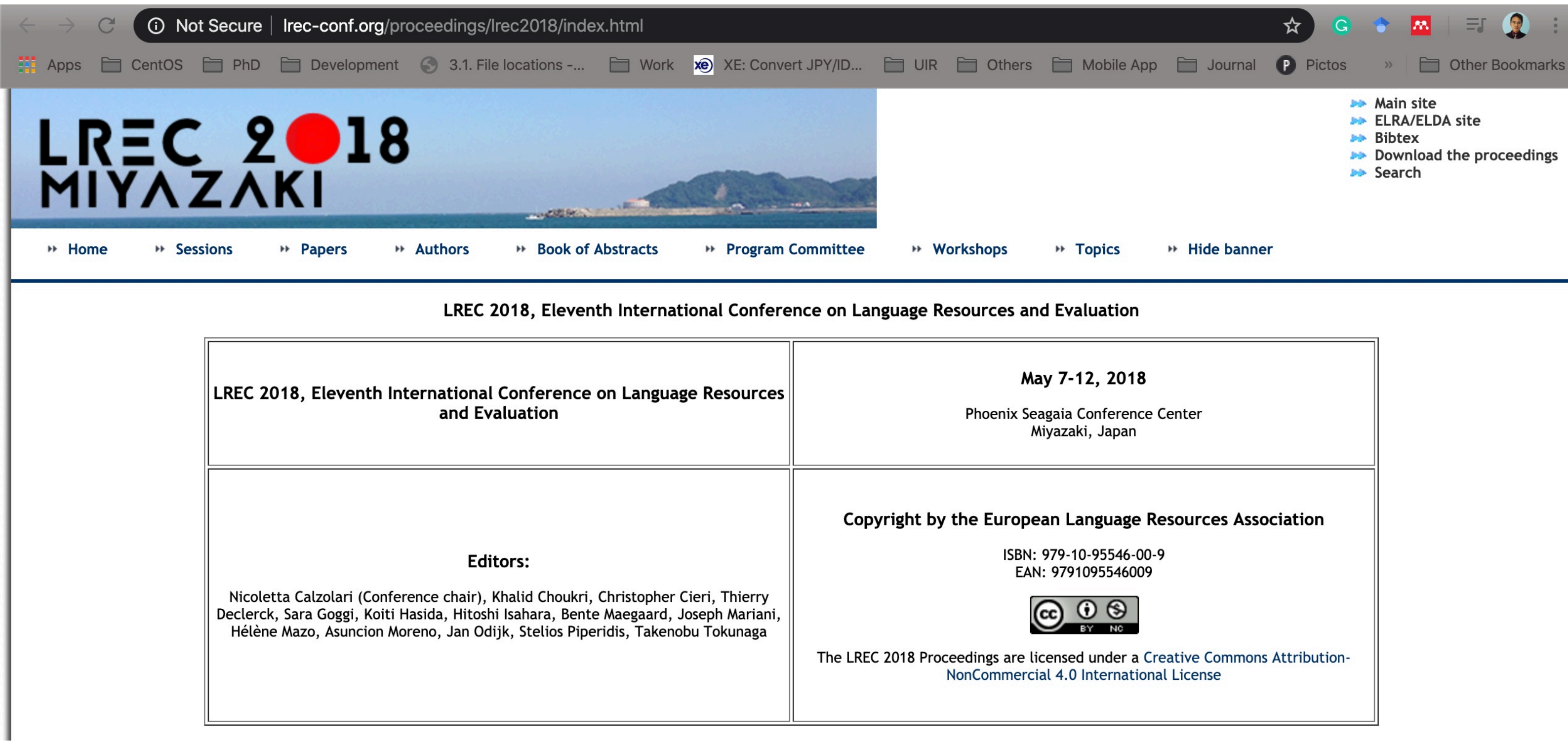
Nasution, A. H., Murakami, Y., and Ishida, T. (2017c). A generalized constraint approach to bilingual dictionary induction for low-resource language families. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 17(2):9:1–9:29, November.

Nasution, A. H., Murakami, Y., and Ishida, T. (2017d). Similarity cluster of indonesian ethnic languages. In *Proceedings of the First International Conference on Science Engineering and Technology (ICoSET 2017)*, pages 12–27, Pekanbaru, Indonesia, November.

Soderland, S., Etzioni, O., Weld, D. S., Skinner, M., Bilmes, J., et al. (2009). Compiling a massive, multilingual dictionary via probabilistic inference. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 262–270. Association for Computational Linguistics.

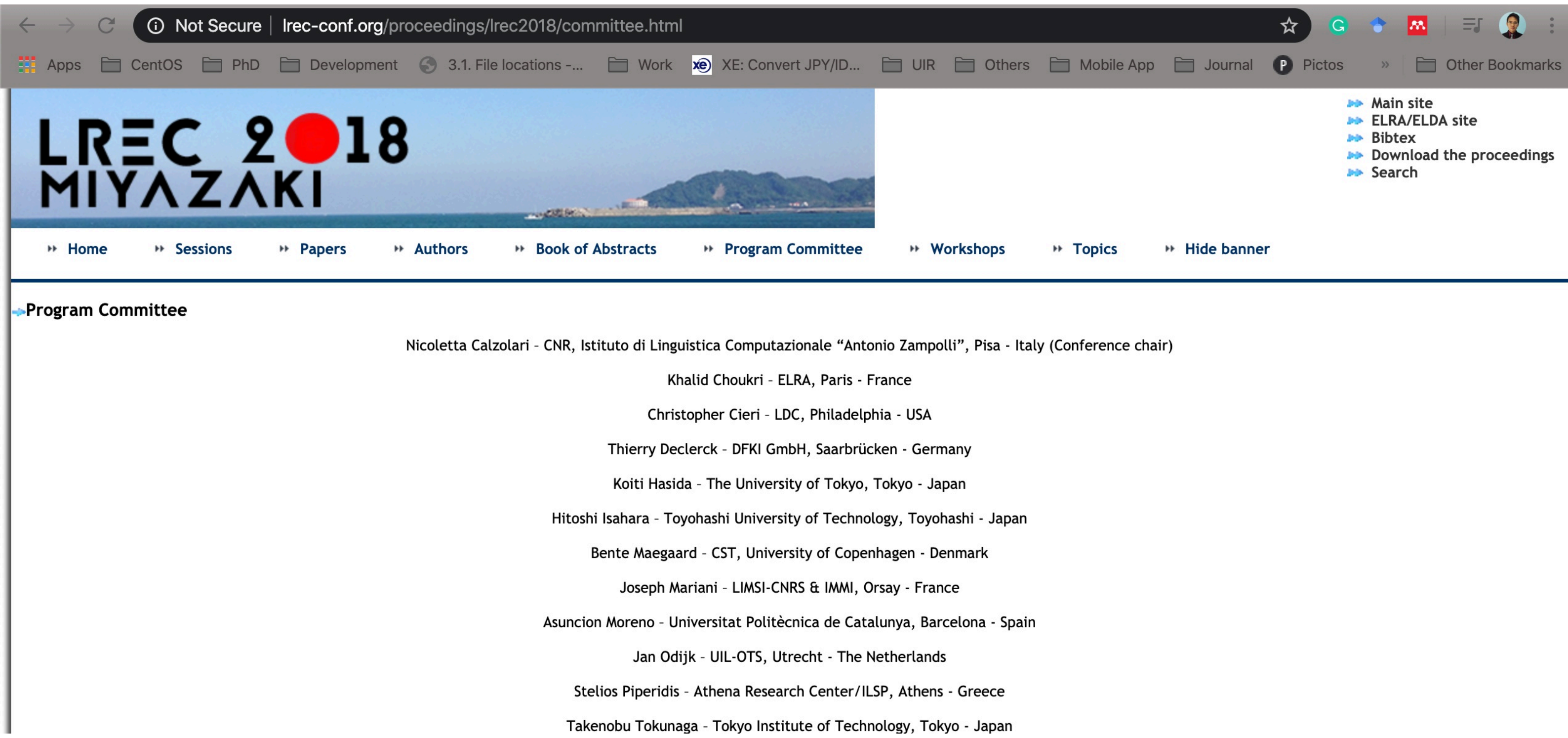
Tanaka, K. and Umemura, K. (1994). Construction of a bilingual dictionary intermediated by a third language. In *Proceedings of the 15th conference on Computational linguistics-Volume 1*, pages 297–303. Association for Computational Linguistics.

Wushouer, M., Lin, D., Ishida, T., and Hirayama, K. (2015). A constraint approach to pivot-based bilingual dictionary induction. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 15(1):4:1–4:26, November.



LREC 2018, Eleventh International Conference on Language Resources and Evaluation

LREC 2018, Eleventh International Conference on Language Resources and Evaluation	**May 7-12, 2018** Phoenix Seagaia Conference Center Miyazaki, Japan
Editors: Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, Takenobu Tokunaga	**Copyright by the European Language Resources Association** ISBN: 979-10-95546-00-9 EAN: 9791095546009 The LREC 2018 Proceedings are licensed under a Creative Commons Attribution-NonCommercial 4.0 International License



Program Committee

Nicoletta Calzolari - CNR, Istituto di Linguistica Computazionale “Antonio Zampolli”, Pisa - Italy (Conference chair)

Khalid Choukri - ELRA, Paris - France

Christopher Cieri - LDC, Philadelphia - USA

Thierry Declerck - DFKI GmbH, Saarbrücken - Germany

Koiti Hasida - The University of Tokyo, Tokyo - Japan

Hitoshi Isahara - Toyohashi University of Technology, Toyohashi - Japan

Bente Maegaard - CST, University of Copenhagen - Denmark

Joseph Mariani - LIMSI-CNRS & IMMI, Orsay - France

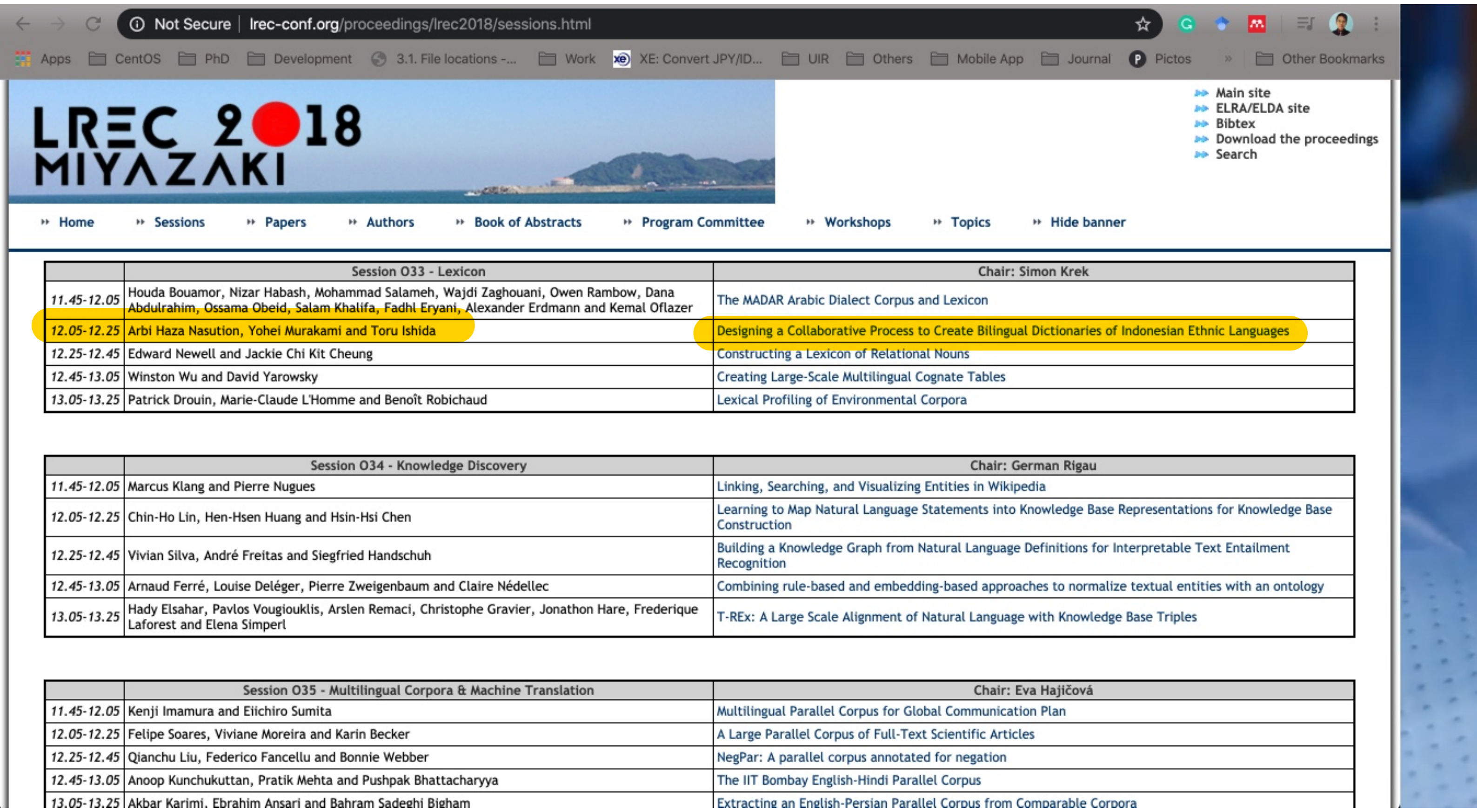
Asuncion Moreno - Universitat Politècnica de Catalunya, Barcelona - Spain

Jan Odijk - UIL-OTS, Utrecht - The Netherlands

Stelios Piperidis - Athena Research Center/ILSP, Athens - Greece

Takenobu Tokunaga - Tokyo Institute of Technology, Tokyo - Japan

Table of Contents



LREC 2018 MIYAZAKI

- Main site
- ELRA/ELDA site
- Bibtex
- Download the proceedings
- Search

Home · Sessions · Papers · Authors · Book of Abstracts · Program Committee · Workshops · Topics · Hide banner

	Session O33 - Lexicon	Chair: Simon Krek
11.45-12.05	Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann and Kemal Oflazer	The MADAR Arabic Dialect Corpus and Lexicon
12.05-12.25	Arbi Haza Nasution, Yohei Murakami and Toru Ishida	Designing a Collaborative Process to Create Bilingual Dictionaries of Indonesian Ethnic Languages
12.25-12.45	Edward Newell and Jackie Chi Kit Cheung	Constructing a Lexicon of Relational Nouns
12.45-13.05	Winston Wu and David Yarowsky	Creating Large-Scale Multilingual Cognate Tables
13.05-13.25	Patrick Drouin, Marie-Claude L'Homme and Benoît Robichaud	Lexical Profiling of Environmental Corpora

	Session O34 - Knowledge Discovery	Chair: German Rigau
11.45-12.05	Marcus Klang and Pierre Nugues	Linking, Searching, and Visualizing Entities in Wikipedia
12.05-12.25	Chin-Ho Lin, Hen-Hsen Huang and Hsin-Hsi Chen	Learning to Map Natural Language Statements into Knowledge Base Representations for Knowledge Base Construction
12.25-12.45	Vivian Silva, André Freitas and Siegfried Handschuh	Building a Knowledge Graph from Natural Language Definitions for Interpretable Text Entailment Recognition
12.45-13.05	Arnaud Ferré, Louise Deléger, Pierre Zweigenbaum and Claire Nédellec	Combining rule-based and embedding-based approaches to normalize textual entities with an ontology
13.05-13.25	Hady Elsahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon Hare, Frederique Laforest and Elena Simperl	T-REx: A Large Scale Alignment of Natural Language with Knowledge Base Triples

	Session O35 - Multilingual Corpora & Machine Translation	Chair: Eva Hajičová
11.45-12.05	Kenji Imamura and Eiichiro Sumita	Multilingual Parallel Corpus for Global Communication Plan
12.05-12.25	Felipe Soares, Viviane Moreira and Karin Becker	A Large Parallel Corpus of Full-Text Scientific Articles
12.25-12.45	Qianchu Liu, Federico Fancellu and Bonnie Webber	NegPar: A parallel corpus annotated for negation
12.45-13.05	Anoop Kunchukuttan, Pratik Mehta and Pushpak Bhattacharyya	The IIT Bombay English-Hindi Parallel Corpus
13.05-13.25	Akbar Karimi, Ebrahim Ansari and Bahram Sadeghi Bigham	Extracting an English-Persian Parallel Corpus from Comparable Corpora