

LREC 2020
Marseille

**TWELFTH INTERNATIONAL CONFERENCE ON
LANGUAGE RESOURCES AND EVALUATION**

May 11-16 , 2020
PALAIS DU PHARO
Marseille, France

CONFERENCE PROCEEDINGS

Editors:

Nicoletta Calzolari (Conference chair), Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis

© The European Language Resources Association (ELRA), 2020

The LREC 2020 Proceedings are licensed under a
Creative Commons Attribution Non-Commercial 4.0 International License

**LREC 2020, TWELFTH INTERNATIONAL CONFERENCE ON
LANGUAGE RESOURCES AND EVALUATION**

Title: LREC 2020 Conference Proceedings

Distributed by

ELRA – European Language Resources Association
9, rue des Cordelières
75013 Paris
France
www.elra.info and www.elda.org
lrec@elda.org

ISBN: 979-10-95546-34-4

EAN: 9791095546344

Introduction to LREC 2020 by Nicoletta Calzolari
Chair of the 12th edition of LREC
ELRA Honorary President

Welcome to the 12th edition of LREC ... that should have been in Marseille, first time in France! Unfortunately not now, in May 2020. Now my welcome is completely virtual, to all of you authors of these Proceedings papers and to the colleagues who will look at these. Virtual but not less sincere.

This LREC would have also been an occasion to celebrate the 25th anniversary of ELRA. We are proud that ELRA is becoming a mature association. And LREC too. LREC started in 1998, 22 years ago.

We hope to welcome you in a non-virtual way next year in Marseille. We will enjoy together not only the conference but also the special “light” of Marseille and the wonderful view of the Mediterranean and the city from the Palais du Pharo.

Some LREC 2020 figures: previous records broken again!

Again all LREC records broken. We received 1468 submissions, plus 18 for the REPROLANG Track. 55 Workshop proposals were submitted, and 14 Tutorials.

We recruited 1715 colleagues to be part of the Scientific Committee. 3162 were invited, out of which 469 refused and 978 didn't answer. Finding reviewers is becoming a quite difficult task, good reviewers even more.

The LREC Program Committee analysed more than 4300 reviews, to reach a decision during a face-to-face meeting. Despite all our efforts, we know that both the reviewing and the decision (meta-reviewing) phases are exposed to various possible problems, from subjectivity to actual mistakes, and we apologise for these.

As a result of reviewing and meta-reviewing, we have 883 papers accepted, of which 188 Orals and 695 Posters. The REPROLANG Committee accepted 12 papers. In addition, we accepted 38 Workshops and 7 Tutorials.

If this had been a normal LREC it would have been for sure the largest ever, but here I cannot put any figure, just imagine ...

LREC 2020 at the time of COVID-19

All this happened before COVID. Now LREC does not take place in May 2020. When we, the ELRA Board and the LREC Program Committee, saw that it would not have been possible to hold LREC in May, we decided not to turn LREC in a virtual conference, differently from what many other conferences are doing this year.

We all know that the value of LREC is not only in listening to the papers, but in the networking, the face-to-face interactions in the crowded poster sessions, the exchange of ideas, also the gathering in the social events (exactly all what we cannot do now). Thinking at the wonderful LREC Poster sessions these days looks like a nightmare, with all the people crowded around the posters. Many things are changing and our perceptions are different these days. It may happen that we have to rethink the way we are used to organise large conferences.

We hope and will do our best to be able to organise LREC 2020 ... maybe next year, in the same place, Marseille, possibly in May.

If we succeed to have LREC next year, as we hope, we will see if to have a partially different model, for example with a mixture of current and new submissions. But we will consider different possibilities.

LREC 2020 Proceedings

So what happens now? We decided that we must publish the Proceedings, here they are. And we asked also all the Workshop organisers to publish all the Workshop Proceedings. Almost all did that.

Your efforts as authors and reviewers will not be lost.

Given that we don't have a Programme, the main Conference Proceedings contain all the accepted papers together, Orals and Posters. This is in line with what we have always said: that the distinction is not based on the value of the paper but is based on their being better suited for one or the other mode of presentation.

The Proceedings are clearly not organised as usual according to the programme sessions, not having a programme, but we decided to cluster them in virtual sessions, around a number of major topics.

This time we also decided to produce all the Proceedings using START, both the Main Conference and the workshops, so that it is straightforward to have all of them in the ACL Anthology.

LREC 2020 acceptance rate: a motivated choice for an inclusive conference

This year, the LREC acceptance rate is 60%, which is very similar to previous years. We are well aware that this is higher than in many other major conferences. But we think that we need conferences of both types. For LREC, as I always state, this is a carefully reasoned decision. For us it is important to try to provide a comprehensive picture of the field and to show how it is evolving. This implies not only to look for new methodologies but also to cover 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: this is for us a very important feature.

What is interesting is that quality and influence is not undermined by our acceptance rate: in 2019 Google Scholar Metrics h5-index, LREC ranks 4th of the large conferences in Computational Linguistics.

LREC 2020 novelty: the REPROLANG2020 Track

An important novelty this year has been the organisation of the REPROLANG Track, as a collaborative shared task. This followed the success of the Workshops on Replicability and Reproducibility that we organised in the last two LRECs. With the organisation of this track inside the main conference, we wanted to promote the view that replicability/reproducibility of research results must be considered as a regular part of scientific practice. We hope this track becomes a tradition in next LRECs. You find these papers, as well as a detailed description of the Track organisation, in one of the Sessions in the Proceedings.

Some LREC 2020 trend

Just few notes on my perception – completely subjective and also impressionistic – of major topics of this year, some typically covered at LREC more than at other conferences.

Among the areas that are increasing I can mention:

- Less-Resourced Languages: this has always been a very important area at LREC
- Analysis of Social Media
- Dialogue and Discourse
- Knowledge discovery, extraction and classification
- Digital Humanities

Always very relevant are:

- Infrastructural issues: another area characteristic of LREC
- Corpora
- Lexicons
- Multilinguality
- Semantics
- Tools, systems, applications

And new topics:

- Cognitive methods
- Neural language representation models

LREC Proceedings in Thomson Citation Index and SCOPUS

I remind that since 2010 the LREC Proceedings have been accepted for inclusion in CPCI (Thomson Reuters Conference Proceedings Citation Index). From 2018 the LREC Proceedings are also indexed by SCOPUS (Elsevier). These are important achievements, providing a better recognition to all LREC authors and useful in particular for young colleagues.

LREC and LRE Journal (Springer)

I also mention that substantially extended versions of papers selected by reviewers as the most appropriate will be considered for publication in a special issue of the Language Resources and Evaluation Journal published by Springer (a SCI-indexed journal).

Acknowledgments

As usual, I want to thank here all those who have worked for this strange LREC 2020, and will continue to do so for it to become a real LREC, possibly next year.

I express my deepest gratitude to the Programme Committee members, especially for the huge task of selecting the papers, the workshops and tutorials. I wish to thank each of them, the new ones: Frédéric Béchet and Philippe Blache, and obviously the “old” ones: Khalid Choukri, Chris Cieri, Thierry Declerk,

Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk and Stelios Piperidis. A particular thanks goes to Jan Odijk, who has worked with me in the preparation of the Proceedings. And to Khalid Choukri, who has worked for the preparation of the Proceedings in START.

I thank the ELRA Association: LREC is a major service from ELRA to all the community. And this time I especially thank the ELRA Board Members and the President, António Branco: they have been very helpful in the difficult decisions we had to make this year.

I am especially grateful to, and somehow I feel sorry for, Frédéric Béchet and Philippe Blache, chairs of the Local Committee, who started to deal with the many local matters and who will hopefully continue to do that next year.

I thank the two institutions that always dedicate a great effort to LREC: ELDA in Paris and ILC-CNR in Pisa. I am very proud of the work of Sara Goggi and H el ene Mazo, and I thank them for their continuous dedication and competence in managing the many tasks they have in their hands. Without their commitment for many months, LREC would not happen. And I add my thanks to many other colleagues of the two institutions who contributed in different ways, and in particular to Roberto Bartolini, Val erie Mapelli, Caroline Rannaud, Irene Russo, Alexandre Sicard.

I express my gratitude to the Sponsors that have helped with financial support.

Also on behalf of the Program Committee, I praise our impressively large Scientific Committee. They did a great job, most of them on time.

I thank the Workshop and Tutorial organisers, and I feel sorry for them too. Some of them are holding a virtual workshop and I wish them a successful workshop. I hope they will share their experience with us. I hope that others will be with us when we can have again LREC in real life.

The biggest thanks goes to all the LREC authors, who provide the “substance” to LREC, and give us such a broad picture of the field. I hope to see all of them next year.

This time, sadly, I cannot thank the LREC participants. I will be so glad to do this possibly next year. For now I can only thank the virtual participants, the readers of these Proceedings.

Marseille will be a fantastic Mediterranean location for LREC 2020 . . . on a different year.

I cannot welcome you at LREC 2020, but with all the Programme Committee I hope you find interesting and fruitful readings in the LREC 2020 Proceedings.

Enjoy the LREC 2020 Marseille Proceedings!

Pisa, Italy, May 13, 2020

Nicoletta Calzolari

Chair of the 12th International Conference on Language Resources and Evaluation
and ELRA Honorary President

LREC 2020 – 12th International Conference on Language Resources and Evaluation
Opening Address
António Branco, President of ELRA

In the annals of ELRA, the European Language Resources Association, 2020 will remain as a memorable year. For very good and very bad reasons.

The bad reasons first. In the very same day that I should have been using these notes in Marseille to address the participants in the opening session of the 12th edition of LREC, the International Conference on Language Resources and Evaluation, I am writing them at home. If instead, as scheduled, I was now at the auditorium of the beautiful Palais du Pharo, which overlooks the entrance of Marseille old harbour, I would be staring at a chilling scenario of empty chairs in a totally empty room. This would happen not because the participants would be staying outdoors, understandably mesmerized with the breathtaking view over the Mediterranean Sea and the joyful morning sunlight of Spring, but because all my colleagues who had planned to come, contribute, benefit and enjoy the conference also did not come to Marseille. The covid-19 outbreak and the confinement measures to fight it across the world disrupted the normal course of life everywhere. And LREC is no exception.

With a reinforced spirit of solidarity and camaraderie with which I would have welcomed you to the conference, on behalf of ELRA, I am hereby sending to all colleagues my best wishes that everything is fine with you, that you are keeping safe and that everything will turn out for the best when this ordeal will eventually fade away.

We should not forget also the very good reasons to recall 2020: this is the year when ELRA is celebrating its 25th anniversary! This is in itself a joyful sign of the vitality of our community, and also that the association's *raison d'être* keeps being valid and that its operation has been successful and survived the test of time.

The most visible sign of success of ELRA is its flagship event, the LREC conference, with its first edition 22 years ago, in Granada (Spain). Quantitatively, assessed by Google Scholar h5-index impact scores, LREC is in the top four most impactful world conferences of its area; its impact is in line with the impact of the most impactful journal; and it is the most impactful publication venue promoted by an organization based outside the USA. Qualitatively, I believe, it is by far the most relevant venue of its area supporting inclusiveness and diversity, therein comprising of course language diversity, by supporting the goal of advancing the scientific study of the whole human language in all its different empirical manifestations as 7 000 different languages in the world.

The activity of ELRA is by no means limited to the promotion of LREC. The association offers a wide array of other services to support researchers and the advancement of the science and technology of language. Those of you reading these lines who may not be acquainted yet with some of these services are hereby invited to visit ELRA's website and discover how you can benefit from them — and in case you have not already done so, you are also invited to enrol as an individual member of ELRA, like several hundreds of other colleagues already did. But I could not go without noting that these services include the premier European open platform for the mapping, preservation and distribution of scientific resources, which helps to match producers and users of resources of all types — for speech, text and multimodal processing, for research and innovation, with free and commercial licenses, etc. — under a design concept pioneered by our association 25 years ago. This is a repository that keeps striving as one of the very few, if not the only, independent self-sustaining platform of its kind in the world.

But celebrating our 25th anniversary is also and foremost looking and planning into the future. The celebrations program set up for this commemoration reflects this desideratum. Here are some picks: the conference LT4ALL - Language Technologies for All, organized by ELRA a few months ago at the UNESCO headquarters in Paris, in association with the International Year of Indigenous Languages, that aimed at mobilizing language technology to promote linguistic diversity worldwide; the consultation of the ELRA Board with external experts, invited to a special purpose workshop aimed at stimulating ideas and proposals for directions into which the association should progress; and the online questionnaire addressed to and responded by ELRA members, with a similar aim.

Of course, most important moments in the celebration program are planned for LREC, and they will happen when the conference will take place. On the one hand, a special session is aimed at hearing from participants about their views, ideas and suggestions on the future of the association and LREC, and openly discuss how it should progress so that it can even better contribute to their members' research interests and needs, and vice-versa, so that members can even better contribute to an enhanced association. On the other hand, we will also have a big celebration party, much more needed and deserved now than before, when we will meet again, after overcoming the temporary hurdles caused by covid-19.

I close these notes with brief words of gratitude — in her opening session notes, the conference chair will know much better all those who should be thanked.

On behalf of ELRA, my warmest thanks go to all colleagues who submitted papers to LREC 2020 and who trust that the conference organized by the association is a good venue for the publication of their research results.

And I send to the colleagues in the Local Organization Committee in Provence (France), including Philippe Blache and Frédéric Béchet, in the Program Committee across the world, including its chair Nicoletta Calzolari, in ILC-CNR in Pisa (Italy), including Sara Goggi, and in the ELRA office in Paris (France), including H  l  ne Mazo and the ELRA Secretary General, Khalid Choukri, and to all colleagues who helped them, also my warmest thank you.

This is a thank you all as warm as we can imagine that the Spring Sun is shining now, at the moment that our conference should have started, outside the Palais du Pharo over the Mediterranean Sea — and as it will shine when we will meet again.

Lisbon, Portugal, May 13, 2020
Ant  nio Branco
President of ELRA

LREC 2020 – 12th International Conference on Language Resources and Evaluation
Opening Address
ELRA Secretary General and ELDA Chief Executive Officer

Khalid Choukri

Despite the distance and despite the current lockdown (almost everywhere), a great virtual welcome to everyone from the bottom of our hearts. We (the ELRA/ELDA team) hope that all our friends and their beloved ones are safe and in good spirit during these hard circumstances.

During the last few days and while writing this short message, my feelings went from the Provence atmosphere and Pagnol to Gabriel Garcia Marquez and love in times of Cholera; but with the sincere hope that our whole community (and beyond) is doing well.

22 years after the first LREC in Granada and for this first LREC in France, we had a very ambitious program: both to celebrate the 25th anniversary of ELRA and to share with all of you the best of Provence, one of the most beautiful French regions. Reviewing the last 25 years of ELRA activities serving our community was part of the plan.

This opportunity is lost for a while, LREC 2020 and the celebrations will not take place in May 2020.

It is with a very heavy heart that we had postponed LREC but we are confident that we will meet soon face to face with a chance to remember the best memories from previous LRECs.

Usually, the messages conveyed by the Committees take into account the nice atmosphere that we experience after a couple of days altogether to go beyond the written statements... It is difficult to carry the same sentiment in simple written words, but we all hope good days will come back soon and we, very likely and very hopefully, have learned more about caring for each other.

As you will read in all our messages, we do not give up. LREC is a major milestone in our lives and LREC is more than a simple scientific gathering, it is also one of the best moments when our community gets together, to share science but also friendship and memories.

So we are working hard to organize an LREC 2020 and we are confident that we will have the great pleasure to welcome you in Marseille and in person.

Let me share with you some of the key events and activities we carried out since we all met in Miyazaki.

LT4All in partnership with UNESCO

In December 2019, ELRA organized a 2-day event LT4All, Language Technologies for All, as part of the International Year of Indigenous Languages (IYIL2019), initiated by the United Nations. Recognizing that only 2% of the 7000 languages spoken in the world benefit from Language Technologies, the question asked was: what should be done for the remaining 98%, among which many are in danger of digital extinction, if not complete extinction.

With a strong commitment of ELRA, SIGUL, the ELRA-ISCA joint SIG on under-resourced languages, and Joseph Mariani, former ELRA President, now Honorary President, we managed to invite almost 400 participants from all continents and various backgrounds (language policy makers, language activists, and language technologists). Major players from our field, already aware of, and committed to these issues, took an active part in the conference. Over two days, challenges posed by the digital divide and multilingualism, particularly serious for indigenous languages, were reviewed. Then an overview of all the achievements that represent the state of the art in our field were presented to the large audience. Further to the success of the IYIL2019, including LT4All, followed by a plenary meeting

at the UN Headquarters in New York in 2019 and high-level UN meeting in Mexico in 2020, with ELRA contribution to both of which, the United Nations have decided to dedicate the next decade (2020-2032) to indigenous languages. These developments are very important for ELRA that gained visibility from these events, acknowledging its long-time involvement in providing and promoting Language Technology support for all languages.

As part of this involvement, ELRA supports the organization of a dedicated workshop at LREC 2020 by its SIGUL, the 1st Joint SLTU (Spoken Language Technologies for Under-resourced languages) and CCURL (Collaboration and Computing for Under-Resourced Languages) Workshop.

ELRA, part of important European initiatives

ELRA is involved in several major European initiatives dealing with Language Resources and Language Technology.

ELG, **the European Language Grid**, is a very important project started in Europe, a major initiative that targets the deployment of a marketplace for Language Resources and Language Technology. ELRA will bring its experience in handling issues related to language resources sharing and distribution. We do expect ELRA members to have multiple benefits from the ELG platform as it will allow to efficiently access services and resources that helps develop new applications.

ELRA has continued to be involved in the series of ELRC (**European Language Resources Coordination**) activities, including the collection of language data for Machine Translation and the organization of reach-out workshops all over Europe (EC Member States, plus Iceland and Norway). In this framework and with a strong support from the European commission, the ELRC consortium has secured a large number of resources, in particular for Machine Translation activities. With the encouragement of the European Commission, ELRA has released 200+ resources usable immediately, freely and openly, by our community, offering another distribution channel to this important assets.

ELRA, part of an important initiative to help address issues related to COVID-19

ELRA, with a large number of partners, is also taking part in a COVID-19 initiative that aims at producing specific Languages Resources so the community can carry out research activities related to information retrieval, machine translation, summarization, fake news identification and checking, etc. Another dimension of this initiative is the setup of a number of challenges and evaluation campaigns to assess the performance of current technology with respect to all these applications and more. The ultimate goal being to ensure that our community contributes to safeguard the welfare of human beings using Language technologies.

Other ELRA activities

The activities around the Language Resources at ELRA encompass a number of aspects including the management of legal issues. At the moment, the number of Language Resources in the ELRA catalogue amounts to 1380. This includes large sets of Language Resources that ELRA is now distributing as part of agreements signed in 2019 and 2020 with international partners to offer more assets to our members and the community at large.

I would like to emphasize the large number of agreements we signed with data right holders to bring in more resources with a strong focus on free resources.

ELRA also continues to assist or partners and members in all legal analysis and negotiations related to language resources to make sure that we are using resources that have been legally cleared and that don't

infringe any copyright or other right of third parties. We have also continued our activities in producing language resources in partnership with third parties. For instance we took part in the MGB challenge that required production of colloquial Arabic from North Africa. Our objective was to ensure that such resources were packaged and made available to the community at large.

Shared LRs at LREC 2018

The Languages Resources shared in 2018 have been made available from the LREC 2018 website (<https://bit.ly/3dpWdq3>). ELRA has long recognised the importance of sharing Language Resources (LRs) and making them available to the community. Since the 2014 edition of LREC, participants have been offered the possibility to share their LRs (data, tools, web-services, etc.) when submitting a paper, uploading them in a special LREC repository set up by ELRA. This effort of sharing LRs, linked to the LRE Map initiative (<https://lremap.elra.info>) for their description, contributes to creating a common repository where everyone can deposit and share data.

Acknowledgments

Last but not least I would like express my deep thanks to all our partners that continue to support the organization of LREC despite the current situation hoping like us for an upcoming face-to-face gathering. I would like to thank our Sponsors: Google, Amazon, Lenovo, Bertin, Ontotex, Vecsys and Vocapia.

I would like to thank Philippe Blache and Frédéric Béchet, the LREC Local Committee in Marseille for providing support to both the scientific content and the local organization during the preparatory phase for this LREC Edition in Marseille. Their efforts will not be wasted, our plans for an LREC 2020 in Marseille remain a top priority in our schedule for the months to come. 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, Valérie Mapelli, Vincenzo Parrinelli, Caroline Rannaud, and Alexandre Sicard. I share their disappointment but their efforts will materialize when LREC takes place soon.

I look forward to the LREC 2020 face to face conference. In the meantime, I hope that you will enjoy the fruitful workshops that are going virtual, and ELRA is happy to assist in this. I also hope that these proceedings will reflect the creativity of our community.

I hope that each of you is safe, and so is your family and friends.

Paris, France, May 13, 2020

Khalid Choukri

ELRA Secretary General and ELDA Chief Executive Officer

**Message to LREC 2020 participants by Frédéric Béchet and Philippe Blache,
Chairs of the Local Committee**

May 2020, in Marseille. Sea, Science and Sun, the cocktail was perfect! Breathtaking views of the Vieux Port from the conference location, meetings and discussions around a poster, at a coffee table, or on the best rooftop in the port at sunset, everything was ready! We would have loved to give you all that this spring. Unfortunately it won't be possible.

May 2020, everywhere in the world. Here we are now, with our eyes glued to computer screens, last link with the outside world, following the latest news about the pandemic, praising our medical staff and trying to work, exchange thoughts and ideas, and keep in touch with each other. And this forces us to think: how do we keep in touch, how do we converse, how do we bring out new projects, new ideas if we can't actually be together? LREC has always been an opportunity to be together, sharing these unique creative moments by talking directly to each other in front of a poster or at a coffee break, sharing a meal, having a good time at a social event . . .

So where do we go from here? Between concerns about the carbon footprint of big international conferences such as LREC and about the impact of international gatherings on the spread of a pandemic, how can we keep one of the most valuable assets of our scientific community: making people from all over the world socialize, exchange ideas and work together towards a common goal.

We don't yet have the solution. Of course, virtualization is part of the answer, as many events have gone "virtual" nowadays. However, we also know that nothing will replace direct contact and face-to-face meetings, and that visiting colleagues in other parts of the world is the best way to rethink our beliefs and advance science.

We have to consider and imagine a new model for sharing and working together, looking for new ideas to enable events like LREC to continue to be places of scientific and cultural exchange. Guaranteeing, of course, the safety of our community while allowing us to continue to be researchers, not restricted by any border, including intellectual ones.

We really hope that we will be able to offer you this cocktail of Sea, Science and Sun that we had prepared for you, in the beautiful city of Marseille! We'll find a solution, we promise!

Marseille, France, May 13, 2020
Philippe & Fred
Co-chairs of the Local committee

LREC 2020 Committees

- Conference Programme Committee

- Nicoletta Calzolari ILC/CNR, Pisa, Italy (Conference chair)
- Frédéric Béchet LIS-CNRS, Aix-Marseille Université, Marseille- France
- Philippe Blache CNRS & Aix-Marseille Université, Marseille- France
- Khalid Choukri ELRA, Paris, France
- Christopher Cieri LDC, Philadelphia, USA
- Thierry Declerck DFKI GmbH, Saarbrücken, Germany
- 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

- 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 2020 web site (<https://lrec2020.lrec-conf.org/en/>).

- Local Committee

- Frédéric Béchet LIS-CNRS, Aix-Marseille Université, Marseille- France
- Philippe Blache CNRS & Aix-Marseille Université, Marseille- France

- Conference Editorial Committee

- Sara Goggi ILC/CNR, Pisa, Italy (Co-chair)
- Hélène Mazo ELDA/ELRA, Paris, France (Co-chair)

- Organising Committee

- Roberto Bartolini 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)
- Vincenzo Parrinelli ILC/CNR, Pisa, Italy
- Caroline Rannaud ELDA/ELRA, Paris, France
- Irene Russo ILC/CNR, Pisa, Italy
- Alexandre Sicard ELDA/ELRA, Paris, France

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 (Founding)
 - Amazon AWS (Gold)
 - Lenovo (Gold)
 - Bertin (Bronze)
 - Ontotex (Bronze)
 - Vecsys (Bronze)
 - Vocapia (Bronze)

- 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)

Table of Contents

LREC2020 – 12th International Conference on Language Resources and Evaluation – Opening Addresses

Anaphora, Coreference

Neural Mention Detection

Juntao Yu, Bernd Bohnet and Massimo Poesio 1

A Cluster Ranking Model for Full Anaphora Resolution

Juntao Yu, Alexandra Uma and Massimo Poesio 11

Mandarinograd: A Chinese Collection of Winograd Schemas

Timothée Bernard and Ting Han 21

On the Influence of Coreference Resolution on Word Embeddings in Lexical-semantic Evaluation Tasks

Alexander Henlein and Alexander Mehler 27

NoEl: An Annotated Corpus for Noun Ellipsis in English

Payal Khullar, Kushal Majmundar and Manish Shrivastava 34

An Annotated Dataset of Coreference in English Literature

David Bamman, Olivia Lewke and Anya Mansoor 44

GerDraCor-Coref: A Coreference Corpus for Dramatic Texts in German

Janis Pagel and Nils Reiter 55

A Study on Entity Resolution for Email Conversations

Parag Pravin Dakle, Takshak Desai and Dan Moldovan 65

Model-based Annotation of Coreference

Rahul Aralikkatte and Anders Søgaard 74

French Coreference for Spoken and Written Language

Rodrigo Wilkens, Bruno Oberle, Frédéric Landragin and Amalia Todirascu 80

Cross-lingual Zero Pronoun Resolution

Abdulrahman Aloraini and Massimo Poesio 90

Exploiting Cross-Lingual Hints to Discover Event Pronouns

Sharid Loáiciga, Christian Hardmeier and Asad Sayeed 99

MuDoCo: Corpus for Multidomain Coreference Resolution and Referring Expression Generation

Scott Martin, Shivani Poddar and Kartikeya Upasani 104

Cognitive Methods

Affection Driven Neural Networks for Sentiment Analysis

Rong Xiang, Yunfei Long, Mingyu Wan, Jinghang Gu, Qin Lu and Chu-Ren Huang 112

<i>The Alice Datasets: fMRI & EEG Observations of Natural Language Comprehension</i> Shohini Bhattasali, Jonathan Brennan, Wen-Ming Luh, Berta Franzluebbers and John Hale . . .	120
<i>Modelling Narrative Elements in a Short Story: A Study on Annotation Schemes and Guidelines</i> Elena Mikhalkova, Timofei Protasov, Polina Sokolova, Anastasiia Bashmakova and Anastasiia Drozdova	126
<i>Cortical Speech Databases For Deciphering the Articulatory Code</i> Harald Höge	133
<i>ZuCo 2.0: A Dataset of Physiological Recordings During Natural Reading and Annotation</i> Nora Hollenstein, Marius Troendle, Ce Zhang and Nicolas Langer	138
<i>Linguistic, Kinematic and Gaze Information in Task Descriptions: The LKG-Corpus</i> Tim Reinboth, Stephanie Gross, Laura Bishop and Brigitte Krenn	147
<i>The ACQDIV Corpus Database and Aggregation Pipeline</i> Anna Jancso, Steven Moran and Sabine Stoll	156
<i>Providing Semantic Knowledge to a Set of Pictograms for People with Disabilities: a Set of Links between WordNet and Arasaac: Arasaac-WN</i> Didier Schwab, Pauline Trial, Céline Vaschalde, Loïc Vial, Emmanuelle Esperanca-Rodier and Benjamin Lecouteux	166
<i>Orthographic Codes and the Neighborhood Effect: Lessons from Information Theory</i> Stéphan Tulkens, Dominiek Sandra and Walter Daelemans	172
<i>Understanding the Dynamics of Second Language Writing through Keystroke Logging and Complexity Contours</i> Elma Kerz, Fabio Pruneri, Daniel Wiechmann, Yu Qiao and Marcus Ströbel	182
<i>Design of BCCWJ-EEG: Balanced Corpus with Human Electroencephalography</i> Yohei Oseki and Masayuki Asahara	189
<i>Using the RUPLEX Multichannel Corpus in a Pilot fMRI Study on Speech Disfluencies</i> Katerina Smirnova, Nikolay Korotaev, Yana Panikratova, Irina Lebedeva, Ekaterina Pechenkova and Olga Fedorova	195
Collaborative Resource Construction & Crowdsourcing	
<i>Construction of an Evaluation Corpus for Grammatical Error Correction for Learners of Japanese as a Second Language</i> Aomi Koyama, Tomoshige Kiyuna, Kenji Kobayashi, Mio Arai and Mamoru Komachi	204
<i>Effective Crowdsourcing of Multiple Tasks for Comprehensive Knowledge Extraction</i> Sangha Nam, Minh Lee, Donghwan Kim, Kijong Han, Kuntae Kim, Sooji Yoon, Eun-kyung Kim and Key-Sun Choi	212
<i>Developing a Corpus of Indirect Speech Act Schemas</i> Antonio Roque, Alexander Tsuetaki, Vasanth Sarathy and Matthias Scheutz	220
<i>Quality Estimation for Partially Subjective Classification Tasks via Crowdsourcing</i> Yoshinao Sato and Kouki Miyazawa	229

<i>Crowdsourcing in the Development of a Multilingual FrameNet: A Case Study of Korean FrameNet</i> Younggyun Hahm, Youngbin Noh, Ji Yoon Han, Tae Hwan Oh, Hyonsu Choe, Hansaem Kim and Key-Sun Choi	236
<i>Towards a Reliable and Robust Methodology for Crowd-Based Subjective Quality Assessment of Query-Based Extractive Text Summarization</i> Neslihan Iskender, Tim Polzehl and Sebastian Möller	245
<i>A Seed Corpus of Hindu Temples in India</i> Priya Radhakrishnan	254
<i>Do You Believe It Happened? Assessing Chinese Readers' Veridicality Judgments</i> Yu-Yun Chang and Shu-Kai Hsieh	259
<i>Creating Expert Knowledge by Relying on Language Learners: a Generic Approach for Mass-Producing Language Resources by Combining Implicit Crowdsourcing and Language Learning</i> Lionel Nicolas, Verena Lyding, Claudia Borg, Corina Forascu, Karèn Fort, Katerina Zdravkova, Iztok Kosem, Jaka Čibej, Špela Arhar Holdt, Alice Millour, Alexander König, Christos Rodosthenous, Federico Sangati, Umair ul Hassan, Anisia Katinskaia, Anabela Barreiro, Lavinia Aparaschivei and Yaakov HaCohen-Kerner	268
Computer-Assisted Language Learning (CALL)	
<i>MAGPIE: A Large Corpus of Potentially Idiomatic Expressions</i> Hessel Haagsma, Johan Bos and Malvina Nissim	279
<i>CRWIZ: A Framework for Crowdsourcing Real-Time Wizard-of-Oz Dialogues</i> Francisco Javier Chiyah Garcia, José Lopes, Xingkun Liu and Helen Hastie	288
<i>Effort Estimation in Named Entity Tagging Tasks</i> Inês Gomes, Rui Correia, Jorge Ribeiro and João Freitas	298
<i>Using Crowdsourced Exercises for Vocabulary Training to Expand ConceptNet</i> Christos Rodosthenous, Verena Lyding, Federico Sangati, Alexander König, Umair ul Hassan, Lionel Nicolas, Jolita Horbacauskiene, Anisia Katinskaia and Lavinia Aparaschivei	307
<i>Predicting Multidimensional Subjective Ratings of Children' Readings from the Speech Signals for the Automatic Assessment of Fluency</i> Gérard Bailly, Erika Godde, Anne-Laure Piat-Marchand and Marie-Line Bosse	317
<i>Constructing Multimodal Language Learner Texts Using LARA: Experiences with Nine Languages</i> Elham Akhlaghi, Branislav Bédi, Fatih Bektaş, Harald Berthelsen, Matthias Butterweck, Cathy Chua, Catia Cucchiarin, Gülşen Eryiğit, Johanna Gerlach, Hanieh Habibi, Neasa Ní Chiaráin, Manny Rayner, Steinþór Steingrímsson and Helmer Strik	323
<i>A Dataset for Investigating the Impact of Feedback on Student Revision Outcome</i> Ildiko Pilan, John Lee, Chak Yan Yeung and Jonathan Webster	332
<i>Creating Corpora for Research in Feedback Comment Generation</i> Ryo Nagata, Kentaro Inui and Shin'ichiro Ishikawa	340
<i>Using Multilingual Resources to Evaluate CEFRLex for Learner Applications</i> Johannes Graën, David Alfter and Gerold Schneider	346

<i>Immersive Language Exploration with Object Recognition and Augmented Reality</i>	
Benny Platte, Anett Platte, Christian Roschke, Rico Thomanek, Thony Rolletschke, Frank Zimmer and Marc Ritter	356
<i>A Process-oriented Dataset of Revisions during Writing</i>	
Rianne Conijn, Emily Dux Speltz, Menno van Zaanen, Luuk Van Waes and Evgeny Chukharev-Hudilainen	363
<i>Automated Writing Support Using Deep Linguistic Parsers</i>	
Luís Morgado da Costa, Roger V P Winder, Shu Yun Li, Benedict Christopher Lin Tzer Liang, Joseph Mackinnon and Francis Bond	369
<i>TLT-school: a Corpus of Non Native Children Speech</i>	
Roberto Gretter, Marco Matassoni, Stefano Bannò and Falavigna Daniele	378
<i>Toward a Paradigm Shift in Collection of Learner Corpora</i>	
Anisia Katinskaia, Sardana Ivanova and Roman Yangarber	386
<i>Quality Focused Approach to a Learner Corpus Development</i>	
Roberts Dargis, Ilze Auziņa, Kristīne Levāne-Petrova and Inga Kaija	392
<i>An Exploratory Study into Automated Précis Grading</i>	
Orphee De Clercq and Senne Van Hoecke	397
<i>Adjusting Image Attributes of Localized Regions with Low-level Dialogue</i>	
Tzu-Hsiang Lin, Alexander Rudnicky, Trung Bui, Doo Soon Kim and Jean Oh	405
<i>Alignment Annotation for Clinic Visit Dialogue to Clinical Note Sentence Language Generation</i>	
Wen-wai Yim, Meliha Yetisgen, Jenny Huang and Micah Grossman	413
<i>MultiWOZ 2.1: A Consolidated Multi-Domain Dialogue Dataset with State Corrections and State Tracking Baselines</i>	
Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Goyal, Peter Ku and Dilek Hakkani-Tur	422
<i>A Comparison of Explicit and Implicit Proactive Dialogue Strategies for Conversational Recommendation</i>	
Matthias Kraus, Fabian Fischbach, Pascal Jansen and Wolfgang Minker	429
<i>Conversational Question Answering in Low Resource Scenarios: A Dataset and Case Study for Basque</i>	
Arantxa Otegi, Aitor Agirre, Jon Ander Campos, Aitor Soroa and Eneko Agirre	436
<i>Construction and Analysis of a Multimodal Chat-talk Corpus for Dialog Systems Considering Interpersonal Closeness</i>	
Yoshihiro Yamazaki, Yuya Chiba, Takashi Nose and Akinori Ito	443
<i>BLISS: An Agent for Collecting Spoken Dialogue Data about Health and Well-being</i>	
Jelte van Waterschoot, Iris Hendrickx, Arif Khan, Esther Klabbers, Marcel de Korte, Helmer Strik, Catia Cucchiarini and Mariët Theune	449
<i>The JDCC Corpus: A Large-Scale Multi-Turn Chinese Dialogue Dataset for E-commerce Customer Service</i>	
Meng Chen, Ruixue Liu, Lei Shen, Shaozu Yuan, Jingyan Zhou, Youzheng Wu, Xiaodong He and Bowen Zhou	459

<i>"Cheese!": a Corpus of Face-to-face French Interactions. A Case Study for Analyzing Smiling and Conversational Humor</i>	
Béatrice Priego-Valverde, Brigitte Bigi and Mary Amoyal	467
<i>The Margarita Dialogue Corpus: A Data Set for Time-Offset Interactions and Unstructured Dialogue Systems</i>	
Alberto Chierici, Nizar Habash and Margarita Bicec	476
<i>How Users React to Proactive Voice Assistant Behavior While Driving</i>	
Maria Schmidt, Wolfgang Minker and Steffen Werner	485
<i>Emotional Speech Corpus for Persuasive Dialogue System</i>	
Sara Asai, Koichiro Yoshino, Seitaro Shinagawa, Sakriani Sakti and Satoshi Nakamura	491
<i>Multimodal Analysis of Cohesion in Multi-party Interactions</i>	
Reshmashree Bangalore Kantharaju, Caroline Langlet, Mukesh Barange, Chloé Clavel and Catherine Pelachaud	498
<i>Treating Dialogue Quality Evaluation as an Anomaly Detection Problem</i>	
Rostislav Nedelchev, Ricardo Usbeck and Jens Lehmann	508
<i>Evaluation of Argument Search Approaches in the Context of Argumentative Dialogue Systems</i>	
Niklas Rach, Yuki Matsuda, Johannes Daxenberger, Stefan Ultes, Keiichi Yasumoto and Wolfgang Minker	513
<i>PATE: A Corpus of Temporal Expressions for the In-car Voice Assistant Domain</i>	
Alessandra Zarcone, Touhidul Alam and Zahra Kolagar	523
<i>Mapping the Dialog Act Annotations of the LEGO Corpus into ISO 24617-2 Communicative Functions</i>	
Eugénio Ribeiro, Ricardo Ribeiro and David Martins de Matos	531
<i>Estimating User Communication Styles for Spoken Dialogue Systems</i>	
Juliana Miehle, Isabel Feustel, Julia Hornauer, Wolfgang Minker and Stefan Ultes	540
<i>The ISO Standard for Dialogue Act Annotation, Second Edition</i>	
Harry Bunt, Volha Petukhova, Emer Gilmartin, Catherine Pelachaud, Alex Fang, Simon Keizer and Laurent Prévot	549
<i>The AICO Multimodal Corpus – Data Collection and Preliminary Analyses</i>	
Kristiina Jokinen	559
<i>A Corpus of Controlled Opinionated and Knowledgeable Movie Discussions for Training Neural Conversation Models</i>	
Fabian Galetzka, Chukwuemeka Uchenna Eneh and David Schlangen	565
<i>A French Medical Conversations Corpus Annotated for a Virtual Patient Dialogue System</i>	
Fréjus A. A. Laleye, Gaël de Chalendar, Antonia Blanié, Antoine Brouquet and Dan Behnamou	574
<i>Getting To Know You: User Attribute Extraction from Dialogues</i>	
Chien-Sheng Wu, Andrea Madotto, Zhaojiang Lin, Peng Xu and Pascale Fung	581
<i>Augmenting Small Data to Classify Contextualized Dialogue Acts for Exploratory Visualization</i>	
Abhinav Kumar, Barbara Di Eugenio, Jillian Aurisano and Andrew Johnson	590
<i>RDG-Map: A Multimodal Corpus of Pedagogical Human-Agent Spoken Interactions.</i>	
Maike Paetzel, Deepthi Karkada and Ramesh Manuvinakurike	600

<i>MPDD: A Multi-Party Dialogue Dataset for Analysis of Emotions and Interpersonal Relationships</i> Yi-Ting Chen, Hen-Hsen Huang and Hsin-Hsi Chen	610
<i>“Alexa in the wild” – Collecting Unconstrained Conversations with a Modern Voice Assistant in a Public Environment</i> Ingo Siegert.....	615
<i>EDA: Enriching Emotional Dialogue Acts using an Ensemble of Neural Annotators</i> Chandrakant Bothe, Cornelius Weber, Sven Magg and Stefan Wermter	620
<i>PACO: a Corpus to Analyze the Impact of Common Ground in Spontaneous Face-to-Face Interaction</i> Mary Amoyal, Béatrice Priego-Valverde and Stephane Rauzy	628
<i>Dialogue Act Annotation in a Multimodal Corpus of First Encounter Dialogues</i> Costanza Navarretta and Patrizia Paggio	634
<i>A Conversation-Analytic Annotation of Turn-Taking Behavior in Japanese Multi-Party Conversation and its Preliminary Analysis</i> Mika Enomoto, Yasuharu Den and Yuichi Ishimoto	644
<i>Understanding User Utterances in a Dialog System for Caregiving</i> Yoshihiko Asao, Julien Kloetzer, Junta Mizuno, Dai Saiki, Kazuma Kadowaki and Kentaro Torisawa	653
<i>Designing Multilingual Interactive Agents using Small Dialogue Corpora</i> Donghui Lin, Masayuki Otani, Ryosuke Okuno and Toru Ishida	662
<i>Multimodal Corpus of Bidirectional Conversation of Human-human and Human-robot Interaction during fMRI Scanning</i> Birgit Rauchbauer, Youssef Hmamouche, Brigitte Bigi, Laurent Prévot, Magalie Ochs and Thierry Chaminade	668
<i>The Brain-IHM Dataset: a New Resource for Studying the Brain Basis of Human-Human and Human-Machine Conversations</i> Magalie Ochs, Roxane Bertrand, Aurélie Goujon, Deirdre Bolger, Anne-Sophie Dubarry and Philippe Blache	676
<i>Dialogue-AMR: Abstract Meaning Representation for Dialogue</i> Claire Bonial, Lucia Donatelli, Mitchell Abrams, Stephanie M. Lukin, Stephen Tratz, Matthew Marge, Ron Artstein, David Traum and Clare Voss.....	684
<i>Relation between Degree of Empathy for Narrative Speech and Type of Responsive Utterance in Attentive Listening</i> Koichiro Ito, Masaki Murata, Tomohiro Ohno and Shigeki Matsubara.....	696
<i>Intent Recognition in Doctor-Patient Interviews</i> Robin Rojowiec, Benjamin Roth and Maximilian Fink	702
<i>BrainPredict: a Tool for Predicting and Visualising Local Brain Activity</i> Youssef Hmamouche, Laurent Prévot, Magalie Ochs and Thierry Chaminade	710
<i>MTSI-BERT: A Session-aware Knowledge-based Conversational Agent</i> Matteo Antonio Senese, Giuseppe Rizzo, Mauro Dragoni and Maurizio Morisio	717

<i>Predicting Ratings of Real Dialogue Participants from Artificial Data and Ratings of Human Dialogue Observers</i>	
Kallirroi Georgila, Carla Gordon, Volodymyr Yanov and David Traum	726
<i>Which Model Should We Use for a Real-World Conversational Dialogue System? a Cross-Language Relevance Model or a Deep Neural Net?</i>	
Seyed Hossein Alavi, Anton Leuski and David Traum	735
<i>Chinese Whispers: A Multimodal Dataset for Embodied Language Grounding</i>	
Dimosthenis Kontogiorgos, Elena Sibirtseva and Joakim Gustafson	743
<i>AMUSED: A Multi-Stream Vector Representation Method for Use in Natural Dialogue</i>	
Gaurav Kumar, Rishabh Joshi, Jaspreet Singh and Promod Yenigalla	750
<i>An Annotation Approach for Social and Referential Gaze in Dialogue</i>	
Vidya Somashekarappa, Christine Howes and Asad Sayeed	759
Digital Humanities	
<i>A Penn-style Treebank of Middle Low German</i>	
Hannah Booth, Anne Breitbarth, Aaron Ecaj and Melissa Farasyn	766
<i>Books of Hours. the First Liturgical Data Set for Text Segmentation.</i>	
Amir Hazem, Beatrice Daille, Christopher Kermorvant, Dominique Stutzmann, Marie-Laurence Bonhomme, Martin Maarand and Mélodie Boillet	776
<i>Corpus of Chinese Dynastic Histories: Gender Analysis over Two Millennia</i>	
Sergey Zinin and Yang Xu	785
<i>The Royal Society Corpus 6.0: Providing 300+ Years of Scientific Writing for Humanistic Study</i>	
Stefan Fischer, Jörg Knappen, Katrin Menzel and Elke Teich	794
<i>Corpus REDEWIEDERGABE</i>	
Annelen Brunner, Stefan Engelberg, Fotis Jannidis, Ngoc Duyen Tanja Tu and Lukas Weimer .	803
<i>WeDH - a Friendly Tool for Building Literary Corpora Enriched with Encyclopedic Metadata</i>	
Mattia Egloff and Davide Picca	813
<i>Automatic Section Recognition in Obituaries</i>	
Valentino Sabbatino, Laura Ana Maria Bostan and Roman Klinger	817
<i>SLäNda: An Annotated Corpus of Narrative and Dialogue in Swedish Literary Fiction</i>	
Sara Szymne and Carin Östman	826
<i>RiQuA: A Corpus of Rich Quotation Annotation for English Literary Text</i>	
Sean Papay and Sebastian Padó	835
<i>A Corpus Linguistic Perspective on Contemporary German Pop Lyrics with the Multi-Layer Annotated "Songkorpus"</i>	
Roman Schneider	842
<i>The BDCamões Collection of Portuguese Literary Documents: a Research Resource for Digital Humanities and Language Technology</i>	
Sara Grilo, Márcia Bolrinha, João Silva, Rui Vaz and António Branco	849

<i>Dataset for Temporal Analysis of English-French Cognates</i> Esteban Frossard, Mickael Coustaty, Antoine Doucet, Adam Jatowt and Simon Hengchen	855
<i>Material Philology Meets Digital Onomastic Lexicography: The NordiCon Database of Medieval Nordic Personal Names in Continental Sources</i> Michelle Waldispühl, Dana Dannells and Lars Borin	860
<i>NLP Scholar: A Dataset for Examining the State of NLP Research</i> Saif M. Mohammad	868
<i>The DReaM Corpus: A Multilingual Annotated Corpus of Grammars for the World’s Languages</i> Shafqat Mumtaz Virk, Harald Hammarström, Markus Forsberg and Søren Wichmann	878
<i>LiViTo: Linguistic and Visual Features Tool for Assisted Analysis of Historic Manuscripts</i> Klaus Müller, Aleksej Tikhonov and Roland Meyer	885
<i>TextAnnotator: A UIMA Based Tool for the Simultaneous and Collaborative Annotation of Texts</i> Giuseppe Abrami, Manuel Stoeckel and Alexander Mehler	891
<i>Deduplication of Scholarly Documents using Locality Sensitive Hashing and Word Embeddings</i> Bikash Gyawali, Lucas Anastasiou and Petr Knoth	901
<i>“Voices of the Great War”: A Richly Annotated Corpus of Italian Texts on the First World War</i> Federico Boschetti, irene de felice, Stefano Dei Rossi, Felice Dell’Orletta, Michele Di Giorgio, Martina Miliani, Lucia C. Passaro, Angelica Puddu, Giulia Venturi, Nicola Labanca, Alessandro Lenci and Simonetta Montemagni	911
<i>DEbateNet-mig15: Tracing the 2015 Immigration Debate in Germany Over Time</i> Gabriella Lapesa, Andre Blessing, Nico Blokker, Erenay Dayanik, Sebastian Haunss, Jonas Kuhn and Sebastian Padó	919
<i>A Corpus of Spanish Political Speeches from 1937 to 2019</i> Elena Álvarez-Mellado	928
<i>A New Latin Treebank for Universal Dependencies: Charters between Ancient Latin and Romance Languages</i> Flavio Massimiliano Cecchini, Timo Korhakangas and Marco Passarotti	933
<i>Identification of Indigenous Knowledge Concepts through Semantic Networks, Spelling Tools and Word Embeddings</i> Renato Rocha Souza, Amelie Dorn, Barbara Piringer and Eveline Wandl-Vogt	943
<i>A Multi-Orthography Parallel Corpus of Yiddish Nouns</i> Jonne Saleva	948
<i>An Annotated Corpus of Adjective-Adverb Interfaces in Romance Languages</i> Katharina Gerhalter, Gerlinde Schneider, Christopher Pollin and Martin Hummel	953
<i>Language Resources for Historical Newspapers: the Impresso Collection</i> Maud Ehrmann, Matteo Romanello, Simon Clematide, Phillip Benjamin Ströbel and Raphaël Barman	958
<i>Allgemeine Musikalische Zeitung as a Searchable Online Corpus</i> Bernd Kampe, Tinghui Duan and Udo Hahn	969

<i>Stylometry in a Bilingual Setup</i>	
Silvie Cinkova and Jan Rybicki	977
<i>Dialect Clustering with Character-Based Metrics: in Search of the Boundary of Language and Dialect</i>	
Yo Sato and Kevin Heffernan	985
Discourse Annotation, Representation and Processing	
<i>DiscSense: Automated Semantic Analysis of Discourse Markers</i>	
Damien Sileo, Tim Van de Cruys, Camille Pradel and Philippe Muller	991
<i>ThemePro: A Toolkit for the Analysis of Thematic Progression</i>	
Monica Dominguez, Juan Soler and Leo Wanner	1000
<i>Machine-Aided Annotation for Fine-Grained Proposition Types in Argumentation</i>	
Yohan Jo, Elijah Mayfield, Chris Reed and Eduard Hovy	1008
<i>Chinese Discourse Parsing: Model and Evaluation</i>	
Lin Chuan-An, Shyh-Shiun Hung, Hen-Hsen Huang and Hsin-Hsi Chen	1019
<i>Shallow Discourse Annotation for Chinese TED Talks</i>	
Wanqiu Long, Xinyi Cai, James Reid, Bonnie Webber and Deyi Xiong	1025
<i>The Discussion Tracker Corpus of Collaborative Argumentation</i>	
Christopher Olshefski, Luca Lugini, Ravneet Singh, Diane Litman and Amanda Godley	1033
<i>Shallow Discourse Parsing for Under-Resourced Languages: Combining Machine Translation and Annotation Projection</i>	
Henny Sluyter-Gäthje, Peter Bourgonje and Manfred Stede	1044
<i>A Corpus of Encyclopedia Articles with Logical Forms</i>	
Nathan Rasmussen and William Schuler	1051
<i>The Potsdam Commentary Corpus 2.2: Extending Annotations for Shallow Discourse Parsing</i>	
Peter Bourgonje and Manfred Stede	1061
<i>On the Creation of a Corpus for Coherence Evaluation of Discursive Units</i>	
Elham Mohammadi, Timothe Beiko and Leila Kosseim	1067
<i>Joint Learning of Syntactic Features Helps Discourse Segmentation</i>	
Takshak Desai, Parag Pravin Dakle and Dan Moldovan	1073
<i>Creating a Corpus of Gestures and Predicting the Audience Response based on Gestures in Speeches of Donald Trump</i>	
Verena Ruf and Costanza Navarretta	1081
<i>GeCzLex: Lexicon of Czech and German Anaphoric Connectives</i>	
Lucie Poláková, Kateřina Rysová, Magdaléna Rysová and Jiří Mírovský	1089
<i>DiMLex-Bangla: A Lexicon of Bangla Discourse Connectives</i>	
Debopam Das, Manfred Stede, Soumya Sankar Ghosh and Lahari Chatterjee	1097
<i>Semi-Supervised Tri-Training for Explicit Discourse Argument Expansion</i>	
Rene Knaebel and Manfred Stede	1103

<i>WikiPossessions: Possession Timeline Generation as an Evaluation Benchmark for Machine Reading Comprehension of Long Texts</i>	
Dhivya Chinnappa, Alexis Palmer and Eduardo Blanco	1110
<i>TED-Q: TED Talks and the Questions they Evoke</i>	
Matthijs Westera, Laia Mayol and Hannah Rohde	1118
<i>CzeDLex 0.6 and its Representation in the PML-TQ</i>	
Jiří Mírovský, Lucie Poláková and Pavlína Synková	1128
<i>Corpus for Modeling User Interactions in Online Persuasive Discussions</i>	
Ryo Egawa, Gaku Morio and Katsuhide Fujita	1135
<i>Simplifying Coreference Chains for Dyslexic Children</i>	
Rodrigo Wilkens and Amalia Todirascu	1142
<i>Adapting BERT to Implicit Discourse Relation Classification with a Focus on Discourse Connectives</i>	
Yudai Kishimoto, Yugo Murawaki and Sadao Kurohashi	1152
<i>What Speakers really Mean when they Ask Questions: Classification of Intentions with a Supervised Approach</i>	
Angèle Barbedette and Iris Eshkol-Taravella	1159
<i>Modeling Dialogue in Conversational Cognitive Health Screening Interviews</i>	
Shahla Farzana, Mina Valizadeh and Natalie Parde	1167
<i>Stigma Annotation Scheme and Stigmatized Language Detection in Health-Care Discussions on Social Media</i>	
Nadiya Straton, Hyeju Jang and Raymond Ng	1178
<i>An Annotated Dataset of Discourse Modes in Hindi Stories</i>	
Swapnil Dhanwal, Hritwik Dutta, Hitesh Nankani, Nilay Shrivastava, Yaman Kumar, Junyi Jessy Li, Debanjan Mahata, Rakesh Gosangi, Haimin Zhang, Rajiv Ratn Shah and Amanda Stent	1191
Document Classification, Text categorisation	
<i>Multi-class Multilingual Classification of Wikipedia Articles Using Extended Named Entity Tag Set</i>	
Hassan S. Shavarani and Satoshi Sekine	1197
<i>An Algerian Corpus and an Annotation Platform for Opinion and Emotion Analysis</i>	
Leila Moudjari, Karima Akli-Astouati and Farah Benamara	1202
<i>Transfer Learning from Transformers to Fake News Challenge Stance Detection (FNC-1) Task</i>	
Valeriya Slovikovskaya and Giuseppe Attardi	1211
<i>Scientific Statement Classification over arXiv.org</i>	
Deyan Ginev and Bruce R Miller	1219
<i>Cross-domain Author Gender Classification in Brazilian Portuguese</i>	
Rafael Dias and Ivandré Paraboni	1227
<i>LEDGAR: A Large-Scale Multi-label Corpus for Text Classification of Legal Provisions in Contracts</i>	
Don Tuggener, Pius von Däniken, Thomas Peetz and Mark Cieliebak	1235

<i>Online Near-Duplicate Detection of News Articles</i> Simon Rodier and Dave Carter	1242
<i>Automated Essay Scoring System for Nonnative Japanese Learners</i> Reo Hirao, Mio Arai, Hiroki Shimanaka, Satoru Katsumata and Mamoru Komachi	1250
<i>A Real-World Data Resource of Complex Sensitive Sentences Based on Documents from the Monsanto Trial</i> Jan Neerbek, Morten Eskildsen, Peter Dolog and Ira Assent	1258
<i>Discovering Biased News Articles Leveraging Multiple Human Annotations</i> Konstantina Lazaridou, Alexander Löser, Maria Mestre and Felix Naumann	1268
<i>Corpora and Baselines for Humour Recognition in Portuguese</i> Hugo Gonalo Oliveira, Andr�e Clem�ncio and Ana Alves	1278
<i>FactCorp: A Corpus of Dutch Fact-checks and its Multiple Usages</i> Marten van der Meulen and W. Gudrun Reijniere	1286
<i>Automatic Orality Identification in Historical Texts</i> Katrin Ortmann and Stefanie Dipper	1293
<i>Using Deep Neural Networks with Intra- and Inter-Sentence Context to Classify Suicidal Behaviour</i> Xingyi Song, Johnny Downs, Sumithra Velupillai, Rachel Holden, Maxim Kikoler, Kalina Bontcheva, Rina Dutta and Angus Roberts	1303
<i>A First Dataset for Film Age Appropriateness Investigation</i> Emad Mohamed and Le An Ha	1311
<i>Habibi - a multi Dialect multi National Arabic Song Lyrics Corpus</i> Mahmoud El-Haj	1318
<i>Age Suitability Rating: Predicting the MPAA Rating Based on Movie Dialogues</i> Mahsa Shafaei, Niloofar Safi Samghabadi, Sudipta Kar and Thamar Solorio	1327
<i>Email Classification Incorporating Social Networks and Thread Structure</i> Sakhar Alkhereyf and Owen Rambow	1336
<i>Development and Validation of a Corpus for Machine Humor Comprehension</i> Yuen-Hsien Tseng, Wun-Syuan Wu, Chia-Yueh Chang, Hsueh-Chih Chen and Wei-Lun Hsu	1346
<i>Alector: A Parallel Corpus of Simplified French Texts with Alignments of Misreadings by Poor and Dyslexic Readers</i> N�ria Gala, Ana�s Tack, Ludivine Javourey-Drevet, Thomas Franois and Johannes C. Ziegler	1353
<i>A Corpus for Detecting High-Context Medical Conditions in Intensive Care Patient Notes Focusing on Frequently Readmitted Patients</i> Edward T. Moseley, Joy T. Wu, Jonathan Welt, John Foote, Patrick D. Tyler, David W. Grant, Eric T. Carlson, Sebastian Gehrmann, Franck Dernoncourt and Leo Anthony Celi	1362
<i>Multilingual Stance Detection in Tweets: The Catalonia Independence Corpus</i> Elena Zotova, Rodrigo Agerri, Manuel Nu�ez and German Rigau	1368
<i>An Evaluation of Progressive Neural Networks for Transfer Learning in Natural Language Processing</i> Abdul Moeed, Gerhard Hagerer, Sumit Dugar, Sarthak Gupta, Mainak Ghosh, Hannah Danner, Oliver Mitevski, Andreas Nawroth and Georg Groh	1376

<i>WAC: A Corpus of Wikipedia Conversations for Online Abuse Detection</i> Noé Cécillon, Vincent Labatut, Richard Dufour and Georges Linarès	1382
<i>FloDusTA: Saudi Tweets Dataset for Flood, Dust Storm, and Traffic Accident Events</i> Btool Hamoui, Mourad Mars and Khaled Almotairi	1391
<i>An Annotated Corpus for Sexism Detection in French Tweets</i> Patricia Chiril, Véronique Moriceau, Farah Benamara, Alda Mari, Gloria Origgi and Marlène Coulomb-Gully	1397
<i>Measuring the Impact of Readability Features in Fake News Detection</i> Roney Santos, Gabriela Pedro, Sidney Leal, Oto Vale, Thiago Pardo, Kalina Bontcheva and Carolina Scarton	1404
<i>When Shallow is Good Enough: Automatic Assessment of Conceptual Text Complexity using Shallow Semantic Features</i> Sanja Stajner and Ioana Hulpuş	1414
<i>DecOp: A Multilingual and Multi-domain Corpus For Detecting Deception In Typed Text</i> Pasquale Capuozzo, Ivano Lauriola, Carlo Strapparava, Fabio Aiolli and Giuseppe Sartori ...	1423
<i>Age Recommendation for Texts</i> Alexis Blandin, Gwénoél Lecorvé, Delphine Battistelli and Aline Étienne	1431
<i>Multilingual Twitter Corpus and Baselines for Evaluating Demographic Bias in Hate Speech Recognition</i> Xiaolei Huang, Linzi Xing, Franck Dernoncourt and Michael J. Paul	1440
<i>VICTOR: a Dataset for Brazilian Legal Documents Classification</i> Pedro Henrique Luz de Araujo, Teófilo Emídio de Campos, Fabricio Ataide Braz and Nilton Correia da Silva	1449
<i>Dynamic Classification in Web Archiving Collections</i> Krutarth Patel, Cornelia Caragea and Mark Phillips	1459
<i>Aspect Flow Representation and Audio Inspired Analysis for Texts</i> Larissa Vasconcelos, Claudio Campelo and Caio Jeronimo	1469
<i>Annotating and Analyzing Biased Sentences in News Articles using Crowdsourcing</i> Sora Lim, Adam Jatowt, Michael Färber and Masatoshi Yoshikawa	1478
<i>Evaluation of Deep Gaussian Processes for Text Classification</i> P. Jayashree and P. K. Srijith	1485
Emotion Recognition and Generation	
<i>EmoEvent: A Multilingual Emotion Corpus based on different Events</i> Flor Miriam Plaza del Arco, Carlo Strapparava, L. Alfonso Urena Lopez and Maite Martin ..	1492
<i>MuSE: a Multimodal Dataset of Stressed Emotion</i> Mimansa Jaiswal, Cristian-Paul Bara, Yuanhang Luo, Mihai Burzo, Rada Mihalcea and Emily Mower Provost	1499
<i>Affect inTweets: A Transfer Learning Approach</i> Linrui Zhang, Hsin-Lun Huang, Yang Yu and Dan Moldovan	1511

<i>Annotation of Emotion Carriers in Personal Narratives</i>	
Aniruddha Tammewar, Alessandra Cervone, Eva-Maria Messner and Giuseppe Riccardi	1517
<i>Towards Interactive Annotation for Hesitation in Conversational Speech</i>	
Jane Wottawa, Marie Tahon, Apolline Marin and Nicolas Audibert	1526
<i>Abusive language in Spanish children and young teenager’s conversations: data preparation and short text classification with contextual word embeddings</i>	
Marta R. Costa-jussà, Esther González, Asuncion Moreno and Eudald Cumalat	1533
<i>IIT-H TEMD Semi-Natural Emotional Speech Database from Professional Actors and Non-Actors</i>	
Banothu Rambabu, Kishore Kumar Botsa, Gangamohan Paidi and Suryakanth V Gangashetty	1538
<i>The POTUS Corpus, a Database of Weekly Addresses for the Study of Stance in Politics and Virtual Agents</i>	
Thomas Janssoone, Kévin Bailly, Gaël Richard and Chloé Clavel	1546
<i>GoodNewsEveryone: A Corpus of News Headlines Annotated with Emotions, Semantic Roles, and Reader Perception</i>	
Laura Ana Maria Bostan, Evgeny Kim and Roman Klinger	1554
<i>SOLO: A Corpus of Tweets for Examining the State of Being Alone</i>	
Svetlana Kiritchenko, Will Hipson, Robert Coplan and Saif M. Mohammad	1567
<i>PoKi: A Large Dataset of Poems by Children</i>	
Will Hipson and Saif M. Mohammad	1578
<i>AlloSat: A New Call Center French Corpus for Satisfaction and Frustration Analysis</i>	
Manon Macary, Marie Tahon, Yannick Estève and Anthony Rousseau	1590
<i>Learning the Human Judgment for the Automatic Evaluation of Chatbot</i>	
Shih-Hung Wu and Sheng-Lun Chien	1598
<i>Korean-Specific Emotion Annotation Procedure Using N-Gram-Based Distant Supervision and Korean-Specific-Feature-Based Distant Supervision</i>	
Young-Jun Lee, Chae-Gyun Lim and Ho-Jin Choi	1603
<i>Semi-Automatic Construction and Refinement of an Annotated Corpus for a Deep Learning Framework for Emotion Classification</i>	
Jiajun Xu, Kyosuke Masuda, Hiromitsu Nishizaki, Fumiyo Fukumoto and Yoshimi Suzuki . .	1611
<i>CEASE, a Corpus of Emotion Annotated Suicide notes in English</i>	
Soumitra Ghosh, Asif Ekbal and Pushpak Bhattacharyya	1618
<i>Training a Broad-Coverage German Sentiment Classification Model for Dialog Systems</i>	
Oliver Guhr, Anne-Kathrin Schumann, Frank Bahrmann and Hans Joachim Böhme	1627
<i>An Event-comment Social Media Corpus for Implicit Emotion Analysis</i>	
Sophia Yat Mei Lee and Helena Yan Ping Lau	1633
<i>An Emotional Mess! Deciding on a Framework for Building a Dutch Emotion-Annotated Corpus</i>	
Luna De Bruyne, Orphee De Clercq and Veronique Hoste	1643
<i>PO-EMO: Conceptualization, Annotation, and Modeling of Aesthetic Emotions in German and English Poetry</i>	
Thomas Haider, Steffen Eger, Evgeny Kim, Roman Klinger and Winfried Menninghaus	1652

<i>Learning Word Ratings for Empathy and Distress from Document-Level User Responses</i>	
João Sedoc, Sven Buechel, Yehonathan Nachmany, Anneke Buffone and Lyle Ungar	1664

Evaluation Methodologies

<i>Evaluation of Sentence Representations in Polish</i>	
Slawomir Dadas, Michał Perełkiewicz and Rafał Poświata	1674
<i>Identification of Primary and Collateral Tracks in Stuttered Speech</i>	
Rachid Riad, Anne-Catherine Bachoud-Lévi, Frank Rudzicz and Emmanuel Dupoux	1681
<i>How to Compare Automatically Two Phonological Strings: Application to Intelligibility Measurement in the Case of Atypical Speech</i>	
Alain Ghio, Muriel Lalain, Laurence Giusti, Corinne Fredouille and Virginie Woisard	1689
<i>Evaluating Text Coherence at Sentence and Paragraph Levels</i>	
Sennan Liu, Shuang Zeng and Sujian Li	1695
<i>HardEval: Focusing on Challenging Tokens to Assess Robustness of NER</i>	
Gabriel Bernier-Colborne and Phillippe Langlais	1704
<i>An Evaluation Dataset for Identifying Communicative Functions of Sentences in English Scholarly Papers</i>	
Kenichi Iwatsuki, Florian Boudin and Akiko Aizawa	1712
<i>An Automatic Tool For Language Evaluation</i>	
Fabio Fassetti and Ilaria Fassetti	1721
<i>Which Evaluations Uncover Sense Representations that Actually Make Sense?</i>	
Jordan Boyd-Graber, Fenfei Guo, Leah Findlater and Mohit Iyyer	1727
<i>Diversity, Density, and Homogeneity: Quantitative Characteristic Metrics for Text Collections</i>	
Yi-An Lai, Xuan Zhu, Yi Zhang and Mona Diab	1739
<i>Towards Few-Shot Event Mention Retrieval: An Evaluation Framework and A Siamese Network Approach</i>	
Bonan Min, Yee Seng Chan and Lingjun Zhao	1747
<i>Linguistic Appropriateness and Pedagogic Usefulness of Reading Comprehension Questions</i>	
Andrea Horbach, Itziar Aldabe, Marie Bexte, Oier Lopez de Lacalle and Montse Maritxalar	1753
<i>Dataset Reproducibility and IR Methods in Timeline Summarization</i>	
Leo Born, Maximilian Bacher and Katja Markert	1763
<i>Database Search vs. Information Retrieval: A Novel Method for Studying Natural Language Querying of Semi-Structured Data</i>	
Stefanie Nadig, Martin Braschler and Kurt Stockinger	1772
<i>Why Attention is Not Explanation: Surgical Intervention and Causal Reasoning about Neural Models</i>	
Christopher Grimsley, Elijah Mayfield and Julia R.S. Bursten	1780

<i>Have a Cake and Eat it Too: Assessing Discriminating Performance of an Intelligibility Index Obtained from a Reduced Sample Size</i>	
Anna Marczyk, Alain Ghio, Muriel Lalain, Marie Rebourg, Corinne Fredouille and Virginie Woisard	1791
<i>Evaluation Metrics for Headline Generation Using Deep Pre-Trained Embeddings</i>	
Abdul Moeed, Yang An, Gerhard Hagerer and Georg Groh	1796
<i>LinCE: A Centralized Benchmark for Linguistic Code-switching Evaluation</i>	
Gustavo Aguilar, Sudipta Kar and Thamar Solorio	1803
<i>Paraphrase Generation and Evaluation on Colloquial-Style Sentences</i>	
Eetu Sjöblom, Mathias Creutz and Yves Scherrer	1814
<i>Analyzing Word Embedding Through Structural Equation Modeling</i>	
Namgi Han, Katsuhiko Hayashi and Yusuke Miyao	1823
<i>Evaluation of Lifelong Learning Systems</i>	
Yevhenii Prokopalo, Sylvain Meignier, Olivier Galibert, Loic Barrault and Anthony Larcher .	1833
<i>Interannotator Agreement for Lexico-Semantic Annotation of a Corpus</i>	
Elżbieta Hajnicz	1842
<i>An In-Depth Comparison of 14 Spelling Correction Tools on a Common Benchmark</i>	
Markus Näther	1849
<i>Sentence Level Human Translation Quality Estimation with Attention-based Neural Networks</i>	
Yu Yuan and Serge Sharoff	1858
<i>Evaluating Language Tools for Fifteen EU-official Under-resourced Languages</i>	
Diego Alves, Gaurish Thakkar and Marko Tadić	1866
<i>Word Embedding Evaluation for Sinhala</i>	
Dimuthu Lakmal, Surangika Ranathunga, Saman Peramuna and Indu Herath	1874
<i>Stress Test Evaluation of Transformer-based Models in Natural Language Understanding Tasks</i>	
Carlos Aspillaga, Andrés Carvallo and Vladimir Araujo	1882
<i>Brand-Product Relation Extraction Using Heterogeneous Vector Space Representations</i>	
Arkadiusz Janz, Łukasz Kopociński, Maciej Piasecki and Agnieszka Pluwak	1895
<i>A Tale of Three Parsers: Towards Diagnostic Evaluation for Meaning Representation Parsing</i>	
Maja Buljan, Joakim Nivre, Stephan Oepen and Lilja Øvrelid	1902
Information Extraction, Information Retrieval	
<i>Headword-Oriented Entity Linking: A Special Entity Linking Task with Dataset and Baseline</i>	
Mu Yang, Chi-Yen Chen, Yi-Hui Lee, Qian-hui Zeng, Wei-Yun Ma, Chen-Yang Shih and Wei-Jhih Chen	1910
<i>TableBank: Table Benchmark for Image-based Table Detection and Recognition</i>	
Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, Ming Zhou and Zhoujun Li	1918

<i>WIKIR: A Python Toolkit for Building a Large-scale Wikipedia-based English Information Retrieval Dataset</i>	
Jibril Frej, Didier Schwab and Jean-Pierre Chevallet	1926
<i>Constructing a Public Meeting Corpus</i>	
Koji Tanaka, Chenhui Chu, Haolin Ren, Benjamin Renoust, Yuta Nakashima, Noriko Takemura, Hajime Nagahara and Takao Fujikawa.....	1934
<i>Annotating and Extracting Synthesis Process of All-Solid-State Batteries from Scientific Literature</i>	
Fusataka Kuniyoshi, Kohei Makino, Jun Ozawa and Makoto Miwa.....	1941
<i>WEXEA: Wikipedia EXhaustive Entity Annotation</i>	
Michael Strobl, Amine Trabelsi and Osmar Zaiane.....	1951
<i>Handling Entity Normalization with no Annotated Corpus: Weakly Supervised Methods Based on Distributional Representation and Ontological Information</i>	
Arnaud Ferré, Robert Bossy, Mouhamadou Ba, Louise Deléger, Thomas Lavergne, Pierre Zweigenbaum and Claire Nédellec.....	1959
<i>HBCP Corpus: A New Resource for the Analysis of Behavioural Change Intervention Reports</i>	
Francesca Bonin, Martin Gleize, Ailbhe Finnerty, Candice Moore, Charles Jochim, Emma Norris, Yufang Hou, Alison J. Wright, Debasis Ganguly, Emily Hayes, Silje Zink, Alessandra Pascale, Pol Mac Aonghusa and Susan Michie	1967
<i>Cross-lingual Structure Transfer for Zero-resource Event Extraction</i>	
Di Lu, Ananya Subburathinam, Heng Ji, Jonathan May, Shih-Fu Chang, Avi Sil and Clare Voss.....	1976
<i>Cross-Domain Evaluation of Edge Detection for Biomedical Event Extraction</i>	
Alan Ramponi, Barbara Plank and Rosario Lombardo.....	1982
<i>Semantic Annotation for Improved Safety in Construction Work</i>	
Paul Thompson, Tim Yates, Emrah Inan and Sophia Ananiadou	1990
<i>Social Web Observatory: A Platform and Method for Gathering Knowledge on Entities from Different Textual Sources</i>	
Leonidas Tsekouras, Georgios Petasis, George Giannakopoulos and Aris Kosmopoulos.....	2000
<i>Development of a Corpus Annotated with Medications and their Attributes in Psychiatric Health Records</i>	
Jaya Chaturvedi, Natalia Viani, Jyoti Sanyal, Chloe Tytherleigh, Idil Hasan, Kate Baird, Sumithra Velupillai, Robert Stewart and Angus Roberts	2009
<i>Do not let the history haunt you: Mitigating Compounding Errors in Conversational Question Answering</i>	
Angrosh Mandya, James O' Neill, Danushka Bollegala and Frans Coenen.....	2017
<i>CLEEK: A Chinese Long-text Corpus for Entity Linking</i>	
Weixin Zeng, Xiang Zhao, Jiuyang Tang, Zhen Tan and Xuqian Huang	2026
<i>The Medical Scribe: Corpus Development and Model Performance Analyses</i>	
Izhak Shafran, Nan Du, Linh Tran, Amanda Perry, Lauren Keyes, Mark Knichel, Ashley Domin, Lei Huang, Yu-hui Chen, Gang Li, Mingqiu Wang, Laurent El Shafey, Hagen Soltau and Justin Stuart Paul	2036
<i>A Contract Corpus for Recognizing Rights and Obligations</i>	
Ruka Funaki, Yusuke Nagata, Kohei Suenaga and Shinsuke Mori.....	2045

<i>Recognition of Implicit Geographic Movement in Text</i> Scott Pezanowski and Prasenjit Mitra	2054
<i>Extraction of the Argument Structure of Tokyo Metropolitan Assembly Minutes: Segmentation of Question-and-Answer Sets</i> Keiichi Takamaru, Yasutomo Kimura, Hideyuki Shibuki, Hokuto Ototake, Yuzu Uchida, Kotaro Sakamoto, Madoka Ishioroshi, Teruko Mitamura and Noriko Kando	2064
<i>A Term Extraction Approach to Survey Analysis in Health Care</i> Cécile Robin, Mona Isazad Mashinchi, Fatemeh Ahmadi Zeleti, Adegboyega Ojo and Paul Buitelaar	2069
<i>A Scientific Information Extraction Dataset for Nature Inspired Engineering</i> Ruben Kruiper, Julian F.V. Vincent, Jessica Chen-Burger, Marc P.Y. Desmulliez and Ioannis Konstas	2078
<i>Automated Discovery of Mathematical Definitions in Text</i> Natalia Vanetik, Marina Litvak, Sergey Shevchuk and Lior Reznik	2086
<i>WN-Saliency: A Corpus of News Articles with Entity Saliency Annotations</i> Chuan Wu, Evangelos Kanoulas, Maarten de Rijke and Wei Lu	2095
<i>Event Extraction from Unstructured Amharic Text</i> ephrem tadesse, Rosa Tsegaye and Kuulaa Qaqqabaa	2103
<i>Comparing Machine Learning and Deep Learning Approaches on NLP Tasks for the Italian Language</i> Bernardo Magnini, Alberto Lavelli and Simone Magnolini	2110
<i>MyFixit: An Annotated Dataset, Annotation Tool, and Baseline Methods for Information Extraction from Repair Manuals</i> Nima Nabizadeh, Dorothea Kolossa and Martin Heckmann	2120
<i>Towards Entity Spaces</i> Marieke van Erp and Paul Groth	2129
<i>Love Me, Love Me, Say (and Write!) that You Love Me: Enriching the WASABI Song Corpus with Lyrics Annotations</i> Michael Fell, Elena Cabrio, Elmahdi Korfed, Michel Buffa and Fabien Gandon	2138
<i>Evaluating Information Loss in Temporal Dependency Trees</i> Mustafa Ocal and Mark Finlayson	2148
<i>Populating Legal Ontologies using Semantic Role Labeling</i> Llio Humphreys, Guido Boella, Luigi Di Caro, Livio Robaldo, Leon van der Torre, Sepideh Ghana-vati and Robert Muthuri	2157
<i>PST 2.0 – Corpus of Polish Spatial Texts</i> Michał Marcińczuk, Marcin Oleksy and Jan Wiczorek	2167
<i>Natural Language Premise Selection: Finding Supporting Statements for Mathematical Text</i> Deborah Ferreira and André Freitas	2175
<i>Odinson: A Fast Rule-based Information Extraction Framework</i> Marco A. Valenzuela-Escárcega, Gus Hahn-Powell and Dane Bell	2183

<i>The STEM-ECR Dataset: Grounding Scientific Entity References in STEM Scholarly Content to Authoritative Encyclopedic and Lexicographic Sources</i>	
Jennifer D’Souza, Anett Hoppe, Arthur Brack, Mohmad Yaser Jaradeh, Sören Auer and Ralph Ewerth	2192
<i>MathAlign: Linking Formula Identifiers to their Contextual Natural Language Descriptions</i>	
Maria Alexeeva, Rebecca Sharp, Marco A. Valenzuela-Escárcega, Jennifer Kadowaki, Adarsh Pyarelal and Clayton Morrison	2204
<i>Domain Adapted Distant Supervision for Pedagogically Motivated Relation Extraction</i>	
Oscar Sainz, Oier Lopez de Lacalle, Itziar Aldabe and Montse Maritxalar	2213
<i>Temporal Histories of Epidemic Events (THEE): A Case Study in Temporal Annotation for Public Health</i>	
Jingcheng Niu, Victoria Ng, Gerald Penn and Erin E. Rees	2223
<i>Exploiting Citation Knowledge in Personalised Recommendation of Recent Scientific Publications</i>	
Anita Khadka, Iván Cantador and Miriam Fernandez	2231
<i>A Platform for Event Extraction in Hindi</i>	
Sovan Kumar Sahoo, Saumajit Saha, Asif Ekbal and Pushpak Bhattacharyya	2241
<i>Rad-SpatialNet: A Frame-based Resource for Fine-Grained Spatial Relations in Radiology Reports</i>	
Surabhi Datta, Morgan Ulinski, Jordan Godfrey-Stovall, Shekhar Khanpara, Roy F. Riascos-Castaneda and Kirk Roberts	2251
<i>NLP Analytics in Finance with DoRe: A French 250M Tokens Corpus of Corporate Annual Reports</i>	
Corentin Masson and Patrick Paroubek	2261
<i>The Language of Brain Signals: Natural Language Processing of Electroencephalography Reports</i>	
Ramon Maldonado and Sanda Harabagiu	2268
Knowledge Discovery and Representation	
<i>Humans Keep It One Hundred: an Overview of AI Journey</i>	
Tatiana Shavrina, Anton Emelyanov, Alena Fenogenova, Vadim Fomin, Vladislav Mikhailov, Andrey Evlampiev, Valentin Malykh, Vladimir Larin, Alex Natekin, Aleksandr Vatulin, Peter Romov, Daniil Anastasiev, Nikolai Zinov and Andrey Chertok	2276
<i>Towards Data-driven Ontologies: a Filtering Approach using Keywords and Natural Language Constructs</i>	
Maaïke de Boer and Jack P. C. Verhoosel	2285
<i>A French Corpus and Annotation Schema for Named Entity Recognition and Relation Extraction of Financial News</i>	
Ali Jabbari, Olivier Sauvage, Hamada Zeine and Hamza Chergui	2293
<i>Inferences for Lexical Semantic Resource Building with Less Supervision</i>	
Nadia Bebeskina and Mathieu Lafourcade	2300
<i>Acquiring Social Knowledge about Personality and Driving-related Behavior</i>	
Ritsuko Iwai, Daisuke Kawahara, Takatsune Kumada and Sadao Kurohashi	2306

<i>Implicit Knowledge in Argumentative Texts: An Annotated Corpus</i>	
Maria Becker, Katharina Korfhage and Anette Frank	2316
<i>Multiple Knowledge GraphDB (MKGDB)</i>	
Stefano Faralli, Paola Velardi and Farid Yusifli	2325
<i>Orchestrating NLP Services for the Legal Domain</i>	
Julian Moreno-Schneider, Georg Rehm, Elena Montiel-Ponsoda, Víctor Rodríguez-Doncel, Artem Revenko, Sotirios Karamatakis, Maria Khvalchik, Christian Sageder, Jorge Gracia and Filippo Maganza	2332
<i>Evaluation Dataset and Methodology for Extracting Application-Specific Taxonomies from the Wikipedia Knowledge Graph</i>	
Georgeta Bordea, Stefano Faralli, Fleur Mougin, Paul Buitelaar and Gayo Diallo	2341
<i>Subjective Evaluation of Comprehensibility in Movie Interactions</i>	
Estelle Randria, Lionel Fontan, Maxime Le Coz, Isabelle Ferrané and Julien Pinquier	2348
<i>Representing Multiword Term Variation in a Terminological Knowledge Base: a Corpus-Based Study</i>	
Pilar León-Araúz, Arianne Reimerink and Melania Cabezas-García	2358
<i>Understanding Spatial Relations through Multiple Modalities</i>	
Soham Dan, Hangfeng He and Dan Roth	2368
<i>A Topic-Aligned Multilingual Corpus of Wikipedia Articles for Studying Information Asymmetry in Low Resource Languages</i>	
Dwaipayan Roy, Sumit Bhatia and Prateek Jain	2373
<i>Pártélet: A Hungarian Corpus of Propaganda Texts from the Hungarian Socialist Era</i>	
Zoltán Kmetty, Veronika Vincze, Dorottya Demszky, Orsolya Ring, Balázs Nagy and Martina Katalin Szabó	2381
<i>KORE 50⁺DYWC: An Evaluation Data Set for Entity Linking Based on DBpedia, YAGO, Wikidata, and Crunchbase</i>	
Kristian Noullet, Rico Mix and Michael Färber	2389
<i>Eye4Ref: A Multimodal Eye Movement Dataset of Referentially Complex Situations</i>	
Özge Alacam, Eugen Ruppert, Amr Rekaby Salama, Tobias Staron and Wolfgang Menzel ...	2396
Language Modelling	
<i>SiBert: Enhanced Chinese Pre-trained Language Model with Sentence Insertion</i>	
Jiahao Chen, Chenjie Cao and Xiuyan Jiang	2405
<i>Processing South Asian Languages Written in the Latin Script: the Dakshina Dataset</i>	
Brian Roark, Lawrence Wolf-Sonkin, Christo Kirov, Sabrina J. Mielke, Cibu Johny, Isin Demirsahin and Keith Hall	2413
<i>GM-RKB WikiText Error Correction Task and Baselines</i>	
Gabor Melli, Abdelrhman Eldallal, Bassim Lazem and Olga Moreira	2424
<i>Embedding Space Correlation as a Measure of Domain Similarity</i>	
Anne Beyer, Göran Kauermann and Hinrich Schütze	2431

<i>Wiki-40B: Multilingual Language Model Dataset</i>	
Mandy Guo, Zihang Dai, Denny Vrandečić and Rami Al-Rfou	2440
<i>Know thy Corpus! Robust Methods for Digital Curation of Web corpora</i>	
Serge Sharoff	2453
<i>Evaluating Approaches to Personalizing Language Models</i>	
Milton King and Paul Cook	2461
<i>Class-based LSTM Russian Language Model with Linguistic Information</i>	
Irina Kipyatkova and Alexey Karpov	2470
<i>Adaptation of Deep Bidirectional Transformers for Afrikaans Language</i>	
Sello Ralethe	2475
<i>FlauBERT: Unsupervised Language Model Pre-training for French</i>	
Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier and Didier Schwab	2479
<i>Accelerated High-Quality Mutual-Information Based Word Clustering</i>	
Manuel R. Ciosici, Ira Assent and Leon Derczynski	2491
<i>Rhythmic Proximity Between Natives And Learners Of French - Evaluation of a metric based on the CEFC corpus</i>	
Sylvain Coulange and Solange Rossato	2497
<i>From Linguistic Resources to Ontology-Aware Terminologies: Minding the Representation Gap</i>	
Giulia Speranza, Maria Pia di Buono, Johanna Monti and Federico Sangati	2503
<i>Modeling Factual Claims with Semantic Frames</i>	
Fatma Arslan, Josue Caraballo, Damian Jimenez and Chengkai Li	2511
Less-Resourced and Endangered Languages	
<i>Automatic Transcription Challenges for Inuktitut, a Low-Resource Polysynthetic Language</i>	
Vishwa Gupta and Gilles Boulianne	2521
<i>Geographically-Balanced Gigaword Corpora for 50 Language Varieties</i>	
Jonathan Dunn and Ben Adams	2528
<i>Data Augmentation using Machine Translation for Fake News Detection in the Urdu Language</i>	
Maaz Amjad, Grigori Sidorov and Alisa Zhila	2537
<i>Evaluation of Greek Word Embeddings</i>	
Stamatis Outsios, Christos Karatsalos, Konstantinos Skianis and Michalis Vazirgiannis	2543
<i>A Dataset of Mycenaean Linear B Sequences</i>	
Katerina Papavassiliou, Gareth Owens and Dimitrios Kosmopoulos	2552
<i>The Nunavut Hansard Inuktitut–English Parallel Corpus 3.0 with Preliminary Machine Translation Results</i>	
Eric Joanis, Rebecca Knowles, Roland Kuhn, Samuel Larkin, Patrick Littell, Chi-kiu Lo, Darlene Stewart and Jeffrey Micher	2562

<i>Exploring Bilingual Word Embeddings for Hiligaynon, a Low-Resource Language</i> Leah Michel, Viktor Hangya and Alexander Fraser	2573
<i>A Finite-State Morphological Analyser for Evenki</i> Anna Zueva, Anastasia Kuznetsova and Francis Tyers	2581
<i>Morphology-rich Alphasyllabary Embeddings</i> Amanuel Mersha and Stephen Wu	2590
<i>Localization of Fake News Detection via Multitask Transfer Learning</i> Jan Christian Blaise Cruz, Julianne Agatha Tan and Charibeth Cheng	2596
<i>Evaluating Sentence Segmentation in Different Datasets of Neuropsychological Language Tests in Brazilian Portuguese</i> Edresson Casanova, Marcos Treviso, Lilian Hübner and Sandra Aluísio	2605
<i>Jejueo Datasets for Machine Translation and Speech Synthesis</i> Kyubong Park, Yo Joong Choe and Jiyeon Ham	2615
<i>Speech Corpus of Ainu Folklore and End-to-end Speech Recognition for Ainu Language</i> Kohei Matsuura, Sei Ueno, Masato Mimura, Shinsuke Sakai and Tatsuya Kawahara	2622
<i>Development of a Guarani - Spanish Parallel Corpus</i> Luis Chiruzzo, Pedro Amarilla, Adolfo Ríos and Gustavo Giménez Lugo	2629
<i>AR-ASAG An ARabic Dataset for Automatic Short Answer Grading Evaluation</i> Leila Ouahrani and Djamel Bennouar	2634
<i>Processing Language Resources of Under-Resourced and Endangered Languages for the Generation of Augmentative Alternative Communication Boards</i> Anne Ferger	2644
<i>The Nisvai Corpus of Oral Narrative Practices from Malekula (Vanuatu) and its Associated Language Resources</i> Jocelyn Aznar and Núria Gala	2649
<i>Building a Time-Aligned Cross-Linguistic Reference Corpus from Language Documentation Data (DoReCo)</i> Ludger Paschen, François Delafontaine, Christoph Draxler, Susanne Fuchs, Matthew Stave and Frank Seifart	2657
<i>Benchmarking Neural and Statistical Machine Translation on Low-Resource African Languages</i> Kevin Duh, Paul McNamee, Matt Post and Brian Thompson	2667
<i>Improved Finite-State Morphological Analysis for St. Lawrence Island Yupik Using Paradigm Function Morphology</i> Emily Chen, Hyunji Hayley Park and Lane Schwartz	2676
<i>Towards a Spell Checker for Zamboanga Chavacano Orthography</i> Marcelo Yuji Himoro and Antonio Pareja-Lora	2685
<i>Identifying Sentiments in Algerian Code-switched User-generated Comments</i> Wafia Adouane, Samia Touileb and Jean-Philippe Bernardy	2698
<i>Automatic Creation of Text Corpora for Low-Resource Languages from the Internet: The Case of Swiss German</i> Lucy Linder, Michael Jungo, Jean Hennebert, Claudiu Cristian Musat and Andreas Fischer ..	2706

<i>Evaluating Sub-word Embeddings in Cross-lingual Models</i> Ali Hakimi Parizi and Paul Cook	2712
<i>A Swiss German Dictionary: Variation in Speech and Writing</i> Larissa Schmidt, Lucy Linder, Sandra Djambazovska, Alexandros Lazaridis, Tanja Samardžić and Claudiu Musat	2720
<i>Towards a Corsican Basic Language Resource Kit</i> Laurent Kevers and Stella Retali-Medori	2726
<i>Evaluating the Impact of Sub-word Information and Cross-lingual Word Embeddings on Mi'kmaq Lan- guage Modelling</i> Jeremie Boudreau, Akankshya Patra, Ashima Suvarna and Paul Cook	2736
<i>Exploring a Choctaw Language Corpus with Word Vectors and Minimum Distance Length</i> Jacqueline Brixey, David Sides, Timothy Vizthum, David Traum and Khalil Iskarous	2746
<i>Massive vs. Curated Embeddings for Low-Resourced Languages: the Case of Yorùbá and Twi</i> Jesujoba Alabi, Kwabena Amponsah-Kaakyire, David Adelani and Cristina España-Bonet ...	2754
<i>TROPBank: Turkish PropBank V2.0</i> Neslihan Kara, Deniz Baran Aslan, Büşra Marşan, Özge Bakay, Koray Ak and Olcay Taner Yıldız	2763
<i>Collection and Annotation of the Romanian Legal Corpus</i> Dan Tufiş, Maria Mitrofan, Vasile Păiș, Radu Ion and Andrei Coman	2773
<i>An Empirical Evaluation of Annotation Practices in Corpora from Language Documentation</i> Kilu von Prince and Sebastian Nordhoff	2778
<i>Annotated Corpus for Sentiment Analysis in Odia Language</i> Gaurav Mohanty, Pruthwik Mishra and Radhika Mamidi	2788
<i>Building a Task-oriented Dialog System for Languages with no Training Data: the Case for Basque</i> Maddalen López de Lacalle, Xabier Saralegi and Iñaki San Vicente	2796
<i>SENCORPUS: A French-Wolof Parallel Corpus</i> Elhadji Mamadou Nguer, Alla Lo, Cheikh M. Bamba Dione, Sileye O. Ba and Moussa Lo ...	2803
<i>A Major Wordnet for a Minority Language: Scottish Gaelic</i> Gábor Bella, Fiona McNeill, Rody Gorman, Caoimhin O Donnaile, Kirsty MacDonald, Yamini Chandrashekar, Abed Alhakim Freihat and Fausto Giunchiglia	2812
<i>Crowdsourcing Speech Data for Low-Resource Languages from Low-Income Workers</i> Basil Abraham, Danish Goel, Divya Siddarth, Kalika Bali, Manu Chopra, Monojit Choudhury, Pratik Joshi, Preethi Jyoti, Sunayana Sitaram and Vivek Seshadri	2819
<i>A Resource for Studying Chatino Verbal Morphology</i> Hilaria Cruz, Antonios Anastasopoulos and Gregory Stump	2827
<i>Learnings from Technological Interventions in a Low Resource Language: A Case-Study on Gondi</i> Devansh Mehta, Sebastin Santy, Ramaravind Kommiya Mothilal, Brij Mohan Lal Srivastava, Alok Sharma, Anurag Shukla, Vishnu Prasad, Venkanna U, Amit Sharma and Kalika Bali	2832

<i>Irony Detection in Persian Language: A Transfer Learning Approach Using Emoji Prediction</i> Preni Golazizian, Behnam Sabeti, Seyed Arad Ashrafi Asli, Zahra Majdabadi, Omid Momenzadeh and reza fahmi	2839
<i>Towards Computational Resource Grammars for Runyankore and Rukiga</i> David Bamutura, Peter Ljunglöf and Peter Nebende	2846
<i>Optimizing Annotation Effort Using Active Learning Strategies: A Sentiment Analysis Case Study in Persian</i> Seyed Arad Ashrafi Asli, Behnam Sabeti, Zahra Majdabadi, Preni Golazizian, reza fahmi and Omid Momenzadeh	2855
<i>BanFakeNews: A Dataset for Detecting Fake News in Bangla</i> Md Zobaer Hossain, Md Ashraful Rahman, Md Saiful Islam and Sudipta Kar	2862
<i>A Resource for Computational Experiments on Mapudungun</i> Mingjun Duan, Carlos Fasola, Sai Krishna Rallabandi, Rodolfo Vega, Antonios Anastasopoulos, Lori Levin and Alan W Black	2872
<i>Automated Parsing of Interlinear Glossed Text from Page Images of Grammatical Descriptions</i> Erich Round, Mark Ellison, Jayden Macklin-Cordes and Sacha Beniamine	2878
<i>The Johns Hopkins University Bible Corpus: 1600+ Tongues for Typological Exploration</i> Arya D. McCarthy, Rachel Wicks, Dylan Lewis, Aaron Mueller, Winston Wu, Oliver Adams, Gar- rett Nicolai, Matt Post and David Yarowsky	2884
<i>Towards Building an Automatic Transcription System for Language Documentation: Experiences from Muyu</i> Alexander Zahrer, Andrej Zgank and Barbara Schuppler	2893
<i>Towards Flexible Cross-Resource Exploitation of Heterogeneous Language Documentation Data</i> Daniel Jettka and Timm Lehmborg	2901
<i>CantoMap: a Hong Kong Cantonese MapTask Corpus</i> Grégoire Winterstein, Carmen Tang and Regine Lai	2906
<i>No Data to Crawl? Monolingual Corpus Creation from PDF Files of Truly low-Resource Languages in Peru</i> Gina Bustamante, Arturo Oncevay and Roberto Zariquiey	2914
<i>Creating a Parallel Icelandic Dependency Treebank from Raw Text to Universal Dependencies</i> Hildur Jónsdóttir and Anton Karl Ingason	2924
<i>Building a Universal Dependencies Treebank for Occitan</i> Aleksandra Miletic, Myriam Bras, Marianne Vergez-Couret, Louise Esher, Clamença Poujade and Jean Sibille	2932
<i>Building the Old Javanese Wordnet</i> David Moeljadi and Zakariya Pamuji Aminullah	2940
<i>CPLM, a Parallel Corpus for Mexican Languages: Development and Interface</i> Gerardo Sierra Martínez, Cynthia Montaña, Gemma Bel-Enguix, Diego Córdova and Margarita Mota Montoya	2947

<i>SiNER: A Large Dataset for Sindhi Named Entity Recognition</i> Wazir Ali, Junyu Lu and Zenglin Xu	2953
---	------

Lexicon, Lexical Database

<i>Construct a Sense-Frame Aligned Predicate Lexicon for Chinese AMR Corpus</i> Li Song, Yuling Dai, Yihuan Liu, Bin Li and Weiguang Qu	2962
<i>MultiMWE: Building a Multi-lingual Multi-Word Expression (MWE) Parallel Corpora</i> Lifeng Han, Gareth Jones and Alan Smeaton	2970
<i>A Myanmar (Burmese)-English Named Entity Transliteration Dictionary</i> Aye Myat Mon, Chenchen Ding, Hour Kaing, Khin Mar Soe, Masao Utiyama and Eiichiro Sumita	2980
<i>CA-EHN: Commonsense Analogy from E-HowNet</i> Peng-Hsuan Li, Tsan-Yu Yang and Wei-Yun Ma	2984
<i>Building Semantic Grams of Human Knowledge</i> Valentina Leone, Giovanni Siragusa, Luigi Di Caro and Roberto Navigli	2991
<i>Automatically Building a Multilingual Lexicon of False Friends With No Supervision</i> Ana Sabina Uban and Liviu P. Dinu	3001
<i>A Parallel WordNet for English, Swedish and Bulgarian</i> Krasimir Angelov	3008
<i>ENGLAWI: From Human- to Machine-Readable Wiktionary</i> Franck Sajous, Basilio Calderone and Nabil Hathout	3016
<i>Opening the Romance Verbal Inflection Dataset 2.0: A CLDF lexicon</i> Sacha Beniamine, Martin Maiden and Erich Round	3027
<i>word2word: A Collection of Bilingual Lexicons for 3,564 Language Pairs</i> Yo Joong Choe, Kyubyong Park and Dongwoo Kim	3036
<i>Introducing Lexical Masks: a New Representation of Lexical Entries for Better Evaluation and Exchange of Lexicons</i> Bruno Cartoni, Daniel Calvelo Aros, Denny Vrandecic and Saran Lertpradit	3046
<i>A Large-Scale Leveled Readability Lexicon for Standard Arabic</i> Muhamed Al Khalil, Nizar Habash and Zhengyang Jiang	3053
<i>Preserving Semantic Information from Old Dictionaries: Linking Senses of the 'Altfranzösisches Wörterbuch' to WordNet</i> Achim Stein	3063
<i>Cifu: a Frequency Lexicon of Hong Kong Cantonese</i> Regine Lai and Grégoire Winterstein	3069
<i>Odi et Amo. Creating, Evaluating and Extending Sentiment Lexicons for Latin.</i> Rachele Sprugnoli, Marco Passarotti, Daniela Corbetta and Andrea Peverelli	3078

<i>WordWars: A Dataset to Examine the Natural Selection of Words</i> Saif M. Mohammad	3087
<i>Challenge Dataset of Cognates and False Friend Pairs from Indian Languages</i> Diptesh Kanojia, Malhar Kulkarni, Pushpak Bhattacharyya and Gholamreza Haffari	3096
<i>Development of a Japanese Personality Dictionary based on Psychological Methods</i> Ritsuko Iwai, Daisuke Kawahara, Takatsune Kumada and Sadao Kurohashi	3103
<i>A Lexicon-Based Approach for Detecting Hedges in Informal Text</i> Jumayel Islam, Lu Xiao and Robert E. Mercer	3109
<i>Word Complexity Estimation for Japanese Lexical Simplification</i> Daiki Nishihara and Tomoyuki Kajiwara	3114
<i>Inducing Universal Semantic Tag Vectors</i> Da Huo and Gerard de Melo	3121
<i>LexiDB: Patterns & Methods for Corpus Linguistic Database Management</i> Matthew Coole, Paul Rayson and John Mariani	3128
<i>Towards a Semi-Automatic Detection of Reflexive and Reciprocal Constructions and Their Representation in a Valency Lexicon</i> Václava Kettnerová, Marketa Lopatkova, Anna Vernerová and Petra Barancikova	3136
<i>Languages Resources for Poorly Endowed Languages : The Case Study of Classical Armenian</i> Chahan Vidal-Gorène and Aliénor Decours-Perez	3145
<i>Constructing Web-Accessible Semantic Role Labels and Frames for Japanese as Additions to the NPCMJ Parsed Corpus</i> Koichi Takeuchi, Alastair Butler, Iku Nagasaki, Takuya Okamura and Prashant Pardeshi	3153
<i>Large-scale Cross-lingual Language Resources for Referencing and Framing</i> Piek Vossen, Filip Ilievski, Marten Postma, Antske Fokkens, Gosse Minnema and Levi Remijnse	3162
<i>Modelling Etymology in LMF/TEI: The Grande Dicionário Houaiss da Língua Portuguesa Dictionary as a Use Case</i> Fahad Khan, Laurent Romary, Ana Salgado, Jack Bowers, Mohamed Khemakhem and Toma Taso- vac	3172
<i>Linking the TUFs Basic Vocabulary to the Open Multilingual Wordnet</i> Francis Bond, Hiroki Nomoto, Luís Morgado da Costa and Arthur Bond	3181
<i>Some Issues with Building a Multilingual Wordnet</i> Francis Bond, Luis Morgado da Costa, Michael Wayne Goodman, John Philip McCrae and Ahti Lohk	3189
<i>Collocations in Russian Lexicography and Russian Collocations Database</i> Maria Khokhlova	3198
<i>Methodological Aspects of Developing and Managing an Etymological Lexical Resource: Introducing EtymDB-2.0</i> Clémentine Fourrier and Benoît Sagot	3207

<i>OFrLex: A Computational Morphological and Syntactic Lexicon for Old French</i> Gaël Guibon and Benoît Sagot	3217
<i>Automatic Reconstruction of Missing Romanian Cognates and Unattested Latin Words</i> Alina Maria Ciobanu, Liviu P. Dinu and Laurentiu Zoicas	3226
<i>A Multilingual Evaluation Dataset for Monolingual Word Sense Alignment</i> Sina Ahmadi, John Philip McCrae, Sanni Nimb, Fahad Khan, Monica Monachini, Bolette Pedersen, Thierry Declerck, Tanja Wissik, Andrea Bellandi, Irene Pisani, Thomas Troelsgård, Sussi Olsen, Simon Krek, Veronika Lipp, Tamás Váradi, László Simon, András Gyorffy, Carole Tiberius, Tanneke Schoonheim, Yifat Ben Moshe, Maya Rudich, Raya Abu Ahmad, Dorielle Lonke, Kira Kovalenko, Margit Langemets, Jelena Kallas, Oksana Dereza, Theodorus Fransen, David Cillessen, David Lindemann, Mikel Alonso, Ana Salgado, José Luis Sancho, Rafael-J. Ureña-Ruiz, Jordi Porta Zamorano, Kiril Simov, Petya Osenova, Zara Kancheva, Ivaylo Radev, Ranka Stanković, Andrej Perdih and Dejan Gabrovsek	3232
<i>A Broad-Coverage Deep Semantic Lexicon for Verbs</i> James Allen, Hannah An, Ritwik Bose, Will de Beaumont and Choh Man Teng	3243
<i>Computational Etymology and Word Emergence</i> Winston Wu and David Yarowsky	3252
<i>A Dataset of Translational Equivalents Built on the Basis of plWordNet-Princeton WordNet Synset Mapping</i> Ewa Rudnicka and Tomasz Naskręć	3260
<i>TRANSLIT: A Large-scale Name Transliteration Resource</i> Fernando Benites, Gilbert François Duivesteijn, Pius von Däniken and Mark Cieliebak	3265
<i>Computing with Subjectivity Lexicons</i> Caio L. M. Jeronimo, Claudio E. C. Campelo, Leandro Balby Marinho, Allan Sales, Adriano Veloso and Roberta Viola	3272
<i>The ACoLi Dictionary Graph</i> Christian Chiarcos, Christian Fäth and Maxim Ionov	3281
LR National, International Projects, Infrastructural, Policy issues	
<i>Resources in Underrepresented Languages: Building a Representative Romanian Corpus</i> Ludmila Midrigan - Ciochina, Victoria Boyd, Lucila Sanchez-Ortega, Diana Malancea_Malac, Doina Midrigan and David P. Corina	3291
<i>World Class Language Technology - Developing a Language Technology Strategy for Danish</i> Sabine Kirchmeier, Bolette Pedersen, Sanni Nimb, Philip Diderichsen and Peter Juel Henriksen	3297
<i>A Corpus for Automatic Readability Assessment and Text Simplification of German</i> Alessia Battisti, Dominik Pfütze, Andreas Säuberli, Marek Kostrzewa and Sarah Ebling	3302
<i>The CLARIN Knowledge Centre for Atypical Communication Expertise</i> Henk van den Heuvel, Nelleke Oostdijk, Caroline Rowland and Paul Trilsbeek	3312

<i>Corpora of Disordered Speech in the Light of the GDPR: Two Use Cases from the DELAD Initiative</i> Henk van den Heuvel, Aleksei Kelli, Katarzyna Klessa and Satu Salaasti	3317
<i>The European Language Technology Landscape in 2020: Language-Centric and Human-Centric AI for Cross-Cultural Communication in Multilingual Europe</i> Georg Rehm, Katrin Marheinecke, Stefanie Hegele, Stelios Piperidis, Kalina Bontcheva, Jan Hajic, Khalid Choukri, Andrejs Vasiljevs, Gerhard Backfried, Christoph Prinz, Jose Manuel Gomez-Perez, Luc Meertens, Paul Lukowicz, Josef van Genabith, Andrea Lösch, Philipp Slusallek, Morten Irgens, Patrick Gatellier, Joachim Köhler, Laure Le Bars, Dimitra Anastasiou, Albina Auksoriūtė, Núria Bel, António Branco, Gerhard Budin, Walter Daelemans, Koenraad De Smedt, Radovan Garabík, Maria Gavriilidou, Dagmar Gromann, Svetla Koeva, Simon Krek, Cvetana Krstev, Krister Lindén, Bernardo Magnini, Jan Odijk, Maciej Ogrodniczuk, Eiríkur Rögnvaldsson, Mike Rosner, Bolette Pedersen, Inguna Skadina, Marko Tadić, Dan Tufiş, Tamás Váradi, Kadri Vider, Andy Way and François Yvon	3322
<i>A Framework for Shared Agreement of Language Tags beyond ISO 639</i> Frances Gillis-Webber and Sabine Tittel	3333
<i>Gigafida 2.0: The Reference Corpus of Written Standard Slovene</i> Simon Krek, Špela Arhar Holdt, Tomaž Erjavec, Jaka Čibej, Andraz Repar, Polona Gantar, Nikola Ljubešić, Iztok Kosem and Kaja Dobrovoljc	3340
<i>Corpus Query Lingua Franca part II: Ontology</i> Stefan Evert, Oleg Harlamov, Philipp Heinrich and Piotr Banski	3346
<i>A CLARIN Transcription Portal for Interview Data</i> Christoph Draxler, Henk van den Heuvel, Arjan van Hessen, Silvia Calamai and Louise Corti	3353
<i>Ellogon Casual Annotation Infrastructure</i> Georgios Petasis and Leonidas Tsekouras	3360
<i>European Language Grid: An Overview</i> Georg Rehm, Maria Berger, Ela Elsholz, Stefanie Hegele, Florian Kintzel, Katrin Marheinecke, Stelios Piperidis, Milto Deligiannis, Dimitris Galanis, Katerina Gkirtzou, Penny Labropoulou, Kalina Bontcheva, David Jones, Ian Roberts, Jan Hajic, Jana Hamrlová, Lukáš Kačena, Khalid Choukri, Victoria Arranz, Andrejs Vasiljevs, Oriens Anvari, Andis Lagzdīņš, Jūlija Meļņika, Gerhard Backfried, Erinc Dikici, Miroslav Janosik, Katja Prinz, Christoph Prinz, Severin Stampfer, Dorothea Thomas-Aniola, Jose Manuel Gomez-Perez, Andres Garcia Silva, Christian Berrío, Ulrich Germann, Steve Renals and Ondrej Klejch	3366
<i>The Competitiveness Analysis of the European Language Technology Market</i> Andrejs Vasiljevs, Inguna Skadina, Indra Samite, Kaspars Kauliņš, Ēriks Ajausks, Jūlija Meļņika and Aivars Bērziņš	3381
<i>Constructing a Bilingual Hadith Corpus Using a Segmentation Tool</i> Shatha Altammami, Eric Atwell and Ammar Alsalka	3390
<i>Facilitating Corpus Usage: Making Icelandic Corpora More Accessible for Researchers and Language Users</i> Steinþór Steingrímsson, Starkaður Barkarson and Gunnar Thor Örnólfsson	3399
<i>Interoperability in an Infrastructure Enabling Multidisciplinary Research: The case of CLARIN</i> Franciska de Jong, Bente Maegaard, Darja Fišer, Dieter van Uytvanck and Andreas Witt	3406

<i>Language Technology Programme for Icelandic 2019-2023</i>	
Anna Nikulásdóttir, Jón Guðnason, Anton Karl Ingason, Hrafn Loftsson, Eiríkur Rögnvaldsson, Einar Freyr Sigurðsson and Steinþór Steingrímsson	3414
<i>Privacy by Design and Language Resources</i>	
Pawel Kamocki and Andreas Witt	3423
<i>Making Metadata Fit for Next Generation Language Technology Platforms: The Metadata Schema of the European Language Grid</i>	
Penny Labropoulou, Katerina Gkirtzou, Maria Gavriilidou, Miltos Deligiannis, Dimitris Galanis, Stelios Piperidis, Georg Rehm, Maria Berger, Valérie Mapelli, Michael Rigault, Victoria Arranz, Khalid Choukri, Gerhard Backfried, Jose Manuel Gomez-Perez and Andres Garcia-Silva	3428
<i>Related Works in the Linguistic Data Consortium Catalog</i>	
Daniel Jaquette, Christopher Cieri and Denise DiPersio	3438
<i>Language Data Sharing in European Public Services – Overcoming Obstacles and Creating Sustainable Data Sharing Infrastructures</i>	
Lilli Smal, Andrea Lösch, Josef van Genabith, Maria Giagkou, Thierry Declerck and Stephan Busemann	3443
<i>A Progress Report on Activities at the Linguistic Data Consortium Benefitting the LREC Community</i>	
Christopher Cieri, James Fiumara, Stephanie Strassel, Jonathan Wright, Denise DiPersio and Mark Liberman	3449
<i>Digital Language Infrastructures – Documenting Language Actors</i>	
Verena Lyding, Alexander König and Monica Pretti	3457
<i>Samrómur: Crowd-sourcing Data Collection for Icelandic Speech Recognition</i>	
David Erik Mollberg, Ólafur Helgi Jónsson, Sunneva Þorsteinsdóttir, Steinþór Steingrímsson, Eydís Huld Magnúsdóttir and Jon Guðnason	3463
Machine Learning	
<i>Semi-supervised Development of ASR Systems for Multilingual Code-switched Speech in Under-resourced Languages</i>	
Astik Biswas, Emre Yilmaz, Febe De Wet, Ewald Van der westhuizen and Thomas Niesler	3468
<i>CLFD: A Novel Vectorization Technique and Its Application in Fake News Detection</i>	
Michail Mersinias, Stergos Afantenos and Georgios Chalkiadakis	3475
<i>SimplifyUR: Unsupervised Lexical Text Simplification for Urdu</i>	
Namoos Hayat Qasmi, Haris Bin Zia, Awais Athar and Agha Ali Raza	3484
<i>Jamo Pair Encoding: Subcharacter Representation-based Extreme Korean Vocabulary Compression for Efficient Subword Tokenization</i>	
Sangwhan Moon and Naoaki Okazaki	3490
<i>Offensive Language and Hate Speech Detection for Danish</i>	
Gudbjartur Ingi Sigurbergsson and Leon Derczynski	3498
<i>Semi-supervised Deep Embedded Clustering with Anomaly Detection for Semantic Frame Induction</i>	
Zheng Xin Yong and Tiago Timponi Torrent	3509

<i>Search Query Language Identification Using Weak Labeling</i> Ritiz Tambi, Ajinkya Kale and Tracy Holloway King	3520
<i>Automated Phonological Transcription of Akkadian Cuneiform Text</i> Aleksi Sahala, Miikka Silfverberg, Antti Arppe and Krister Lindén	3528
<i>COSTRA 1.0: A Dataset of Complex Sentence Transformations</i> Petra Barancikova and Ondřej Bojar	3535
<i>Automatic In-the-wild Dataset Annotation with Deep Generalized Multiple Instance Learning</i> Joana Correia, Isabel Trancoso and Bhiksha Raj	3542
<i>How Much Data Do You Need? About the Creation of a Ground Truth for Black Letter and the Effectiveness of Neural OCR</i> Phillip Benjamin Ströbel, Simon Clematide and Martin Volk	3551
<i>Dirichlet-Smoothed Word Embeddings for Low-Resource Settings</i> Jakob Jungmaier, Nora Kassner and Benjamin Roth	3560
<i>On The Performance of Time-Pooling Strategies for End-to-End Spoken Language Identification</i> Joao Monteiro, Md Jahangir Alam and Tiago Falk	3566
<i>Neural Disambiguation of Lemma and Part of Speech in Morphologically Rich Languages</i> José María Hoya Quecedo, Koppatz Maximilian and Roman Yangarber	3573
<i>Non-Linearity in Mapping Based Cross-Lingual Word Embeddings</i> Jiawei Zhao and Andrew Gilman	3583

Machine Translation

<i>LibriVoxDeEn: A Corpus for German-to-English Speech Translation and German Speech Recognition</i> Benjamin Beilharz, Xin Sun, Sariya Karimova and Stefan Riezler	3590
<i>SEDAR: a Large Scale French-English Financial Domain Parallel Corpus</i> Abbas Ghaddar and Phillippe Langlais	3595
<i>JParaCrawl: A Large Scale Web-Based English-Japanese Parallel Corpus</i> Makoto Morishita, Jun Suzuki and Masaaki Nagata	3603
<i>Neural Machine Translation for Low-Resourced Indian Languages</i> Himanshu Choudhary, Shivansh Rao and Rajesh Rohilla	3610
<i>Content-Equivalent Translated Parallel News Corpus and Extension of Domain Adaptation for NMT</i> Hideya Mino, Hideki Tanaka, Hitoshi Ito, Isao Goto, Ichiro Yamada and Takenobu Tokunaga	3616
<i>NMT and PBSMT Error Analyses in English to Brazilian Portuguese Automatic Translations</i> Helena Caseli and Marcio Inácio	3623
<i>Evaluation Dataset for Zero Pronoun in Japanese to English Translation</i> Sho Shimazu, Sho Takase, Toshiaki Nakazawa and Naoaki Okazaki	3630
<i>Better Together: Modern Methods Plus Traditional Thinking in NP Alignment</i> Ádám Kovács, Judit Ács, Andras Kornai and Gábor Recski	3635

<i>Coursera Corpus Mining and Multistage Fine-Tuning for Improving Lectures Translation</i> Haiyue Song, Raj Dabre, Atsushi Fujita and Sadao Kurohashi	3640
<i>Being Generous with Sub-Words towards Small NMT Children</i> Arne Defauw, Tom Vanallemersch, Koen Van Winckel, Sara Szoc and Joachim Van den Bogaert	3650
<i>Document Sub-structure in Neural Machine Translation</i> Radina Dobрева, Jie Zhou and Rachel Bawden	3657
<i>An Evaluation Benchmark for Testing the Word Sense Disambiguation Capabilities of Machine Translation Systems</i> Alessandro Raganato, Yves Scherrer and Jörg Tiedemann	3668
<i>MEDLINE as a Parallel Corpus: a Survey to Gain Insight on French-, Spanish- and Portuguese-speaking Authors' Abstract Writing Practice</i> Aurélie Névéol, Antonio Jimeno Yepes and Mariana Neves	3676
<i>JASS: Japanese-specific Sequence to Sequence Pre-training for Neural Machine Translation</i> Zhuoyuan Mao, Fabien Cromieres, Raj Dabre, Haiyue Song and Sadao Kurohashi	3683
<i>A Post-Editing Dataset in the Legal Domain: Do we Underestimate Neural Machine Translation Quality?</i> Julia Ive, Lucia Specia, Sara Szoc, Tom Vanallemersch, Joachim Van den Bogaert, Eduardo Farah, Christine Maroti, Artur Ventura and Maxim Khalilov	3692
<i>Linguistically Informed Hindi-English Neural Machine Translation</i> Vikrant Goyal, Pruthwik Mishra and Dipti Misra Sharma	3698
<i>A Test Set for Discourse Translation from Japanese to English</i> Masaaki Nagata and Makoto Morishita	3704
<i>An Analysis of Massively Multilingual Neural Machine Translation for Low-Resource Languages</i> Aaron Mueller, Garrett Nicolai, Arya D. McCarthy, Dylan Lewis, Winston Wu and David Yarowsky	3710
<i>TDDC: Timely Disclosure Documents Corpus</i> Nobushige Doi, Yusuke Oda and Toshiaki Nakazawa	3719
<i>MuST-Cinema: a Speech-to-Subtitles corpus</i> Alina Karakanta, Matteo Negri and Marco Turchi	3727
<i>On Context Span Needed for Machine Translation Evaluation</i> Sheila Castilho, Maja Popović and Andy Way	3735
<i>A Multilingual Parallel Corpora Collection Effort for Indian Languages</i> Shashank Siripragada, Jerin Philip, Vinay P. Namboodiri and C V Jawahar	3743
<i>To Case or not to case: Evaluating Casing Methods for Neural Machine Translation</i> Thierry Etchegoyhen and Harritxu Gete	3752
<i>The MARCELL Legislative Corpus</i> Tamás Váradi, Svetla Koeva, Martin Yamalov, Marko Tadić, Bálint Sass, Bartłomiej Nitoń, Maciej Ogrodniczuk, Piotr Pezik, Verginica Barbu Mititelu, Radu Ion, Elena Irimia, Maria Mitrofan, Vasile Păiș, Dan Tufiș, Radovan Garabík, Simon Krek, Andraz Repar, Matjaž Rihtar and Janez Brank	3761

<i>ParaPat: The Multi-Million Sentences Parallel Corpus of Patents Abstracts</i> Felipe Soares, Mark Stevenson, Diego Bartolome and Anna Zaretskaya	3769
<i>Corpora for Document-Level Neural Machine Translation</i> Siyou Liu and Xiaojun Zhang	3775
<i>OpusTools and Parallel Corpus Diagnostics</i> Mikko Aulamo, Umut Sulubacak, Sami Virpioja and Jörg Tiedemann	3782
<i>Literary Machine Translation under the Magnifying Glass: Assessing the Quality of an NMT-Translated Detective Novel on Document Level</i> Margot Fonteyne, Arda Tezcan and Lieve Macken	3790
<i>Handle with Care: A Case Study in Comparable Corpora Exploitation for Neural Machine Translation</i> Thierry Etcheoghyen and Harritxu Gete	3799
<i>The FISKMÖ Project: Resources and Tools for Finnish-Swedish Machine Translation and Cross-Linguistic Research</i> Jörg Tiedemann, Tommi Nieminen, Mikko Aulamo, Jenna Kanerva, Akseli Leino, Filip Ginter and Niko Papula	3808
<i>Multiword Expression aware Neural Machine Translation</i> Andrea Zaninello and Alexandra Birch	3816
Morphology and Tagging	
<i>An Enhanced Mapping Scheme of the Universal Part-Of-Speech for Korean</i> Myung Hee Kim and Nathalie Colineau	3826
<i>Finite State Machine Pattern-Root Arabic Morphological Generator, Analyzer and Diacritizer</i> Maha Alkhairy, Afshan Jafri and David Smith	3834
<i>An Unsupervised Method for Weighting Finite-state Morphological Analyzers</i> Amr Keleg, Francis Tyers, Nick Howell and Tommi Pirinen	3842
<i>Language-Independent Tokenisation Rivals Language-Specific Tokenisation for Word Similarity Prediction</i> Danushka Bollegala, Ryuichi Kiryo, Kosuke Tsujino and Haruki Yukawa	3851
<i>A Supervised Part-Of-Speech Tagger for the Greek Language of the Social Web</i> Maria Nefeli Nikiforos and Katia Lida Kermanidis	3861
<i>Bag & Tag'em - A New Dutch Stemmer</i> Anne Jonker, Corné de Ruijt and Jornt de Gruijl	3868
<i>Glawinette: a Linguistically Motivated Derivational Description of French Acquired from GLAWI</i> Nabil Hathout, Franck Sajous, Basilio Calderone and Fiammetta Namer	3877
<i>BabyFST - Towards a Finite-State Based Computational Model of Ancient Babylonian</i> Aleksi Sahala, Miikka Silfverberg, Antti Arppe and Krister Lindén	3886
<i>Morphological Analysis and Disambiguation for Gulf Arabic: The Interplay between Resources and Methods</i> Salam Khalifa, Nasser Zalmout and Nizar Habash	3895

<i>Wikinflection Corpus: A (Better) Multilingual, Morpheme-Annotated Inflectional Corpus</i> Eleni Metheniti and Guenter Neumann	3905
<i>Introducing a Large-Scale Dataset for Vietnamese POS Tagging on Conversational Texts</i> Oanh Tran, Tu Pham, Vu Dang and Bang Nguyen.....	3913

Conversational Systems/Dialogue/Chatbots/Human-Robot Interaction

UniMorph 3.0: Universal Morphology

Arya D. McCarthy, Christo Kirov, Matteo Grella, Amrit Nidhi, Patrick Xia, Kyle Gorman, Ekaterina Vylomova, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, Timofey Arkhangelskiy, Nataly Krizhanovsky, Andrew Krizhanovsky, Elena Klyachko, Alexey Sorokin, John Mansfield, Valts Ernštreits, Yuval Pinter, Cassandra L. Jacobs, Ryan Cotterell, Mans Hulden and David Yarowsky	3922
---	------

Building the Spanish-Croatian Parallel Corpus

Bojana Mikelenić and Marko Tadić.....	3932
---------------------------------------	------

DerivBase.Ru: a Derivational Morphology Resource for Russian

Daniil Vodolazsky	3937
-------------------------	------

Morfessor EM+Prune: Improved Subword Segmentation with Expectation Maximization and Pruning

Stig-Arne Grönroos, Sami Virpioja and Mikko Kurimo	3944
--	------

Machine Learning and Deep Neural Network-Based Lemmatization and Morphosyntactic Tagging for Serbian

Ranka Stankovic, Branislava Šandrih, Cvetana Krstev, Miloš Utvić and Mihailo Skoric	3954
---	------

Fine-grained Morphosyntactic Analysis and Generation Tools for More Than One Thousand Languages

Garrett Nicolai, Dylan Lewis, Arya D. McCarthy, Aaron Mueller, Winston Wu and David Yarowsky	3963
--	------

Cairo Student Code-Switch (CSCS) Corpus: An Annotated Egyptian Arabic-English Corpus

Mohamed Balabel, Injy Hamed, Slim Abdennadher, Ngoc Thang Vu and Özlem Çetinoğlu ..	3973
---	------

Getting More Data for Low-resource Morphological Inflection: Language Models and Data Augmentation

Alexey Sorokin.....	3978
---------------------	------

Visual Modeling of Turkish Morphology

Berke Özenç and Ercan Solak	3984
-----------------------------------	------

Kvistur 2.0: a BiLSTM Compound Splitter for Icelandic

Jón Daðason, David Mollberg, Hrafn Loftsson and Kristín Bjarnadóttir	3991
--	------

Morphological Segmentation for Low Resource Languages

Justin Mott, Ann Bies, Stephanie Strassel, Jordan Kodner, Caitlin Richter, Hongzhi Xu and Mitchell Marcus.....	3996
--	------

Multilinguality

<i>CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data</i> Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin and Edouard Grave	4003
<i>On the Robustness of Unsupervised and Semi-supervised Cross-lingual Word Embedding Learning</i> Yerai Doval, Jose Camacho-Collados, Luis Espinosa Anke and Steven Schockaert	4013
<i>Building an English-Chinese Parallel Corpus Annotated with Sub-sentential Translation Techniques</i> Yuming Zhai, Lufei Liu, Xinyi Zhong, Gbariel Illouz and Anne Vilnat	4024
<i>Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection</i> Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajic, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers and Daniel Zeman	4034
<i>EMPAC: an English–Spanish Corpus of Institutional Subtitles</i> Iris Serrat Roozen and José Manuel Martínez Martínez	4044
<i>Cross-Lingual Word Embeddings for Turkic Languages</i> Elmurod Kuriyozov, Yerai Doval and Carlos Gómez-Rodríguez	4054
<i>How Universal are Universal Dependencies? Exploiting Syntax for Multilingual Clause-level Sentiment Detection</i> Hiroshi Kanayama and Ran Iwamoto	4063
<i>Multilingual Culture-Independent Word Analogy Datasets</i> Matej Ulčar, Kristiina Vaik, Jessica Lindström, Milda Dailidėnaitė and Marko Robnik-Šikonja	4074
<i>GeBioToolkit: Automatic Extraction of Gender-Balanced Multilingual Corpus of Wikipedia Biographies</i> Marta R. Costa-jussà, Pau Li Lin and Cristina España-Bonet	4081
<i>SpiCE: A New Open-Access Corpus of Conversational Bilingual Speech in Cantonese and English</i> Khia A. Johnson, Molly Babel, Ivan Fong and Nancy Yiu	4089
<i>Identifying Cognates in English-Dutch and French-Dutch by means of Orthographic Information and Cross-lingual Word Embeddings</i> Els Lefever, Sofie Labat and Pranaydeep Singh	4096
<i>Lexicogrammatical translationese across two targets and competence levels</i> Maria Kunilovskaya and Ekaterina Lapshinova-Koltunski	4102
<i>UniSent: Universal Adaptable Sentiment Lexica for 1000+ Languages</i> Ehsaneddin Asgari, Fabienne Braune, Benjamin Roth, Christoph Ringlstetter and Mohammad Mofrad	4113
<i>CanVEC - the Canberra Vietnamese-English Code-switching Natural Speech Corpus</i> Li Nguyen and Christopher Bryant	4121
<i>A Spelling Correction Corpus for Multiple Arabic Dialects</i> Fadhl Eryani, Nizar Habash, Houda Bouamor and Salam Khalifa	4130
<i>A Dataset for Multi-lingual Epidemiological Event Extraction</i> Stephen Mutuvi, Antoine Doucet, Gael Lejeune and Moses Odeo	4139

<i>Swiss-AL: A Multilingual Swiss Web Corpus for Applied Linguistics</i>	
Julia Krasselt, Philipp Dressen, Matthias Fluor, Cerstin Mahlow, Klaus Rothenhäusler and Maren Runte	4145
<i>Analysis of GlobalPhone and Ethiopian Languages Speech Corpora for Multilingual ASR</i>	
Martha Yifiru Tachbelie, Solomon Teferra Abate and Tanja Schultz	4152
<i>Multilingualization of Medical Terminology: Semantic and Structural Embedding Approaches</i>	
Long-Huei Chen and Kyo Kageura	4157
<i>Large Vocabulary Read Speech Corpora for Four Ethiopian Languages: Amharic, Tigrigna, Oromo and Wolaytta</i>	
Solomon Teferra Abate, Martha Yifiru Tachbelie, Michael Melese, Hafte Abera, Tewodros Abebe, Wondwossen Mulugeta, Yaregal Assabie, Million Meshesha, Solomon Afnafu and Binyam Ephrem Seyoum	4167
<i>Incorporating Politeness across Languages in Customer Care Responses: Towards building a Multilingual Empathetic Dialogue Agent</i>	
Mauajama Firdaus, Asif Ekbal and Pushpak Bhattacharyya	4172
<i>WikiBank: Using Wikidata to Improve Multilingual Frame-Semantic Parsing</i>	
Cezar Sas, Meriem Beloucif and Anders Søgaard	4183
<i>Multilingual Corpus Creation for Multilingual Semantic Similarity Task</i>	
Mahtab Ahmed, Chahna Dixit, Robert E. Mercer, Atif Khan, Muhammad Rifayat Samee and Felipe Urra	4190
<i>CoVoST: A Diverse Multilingual Speech-To-Text Translation Corpus</i>	
Changhan Wang, Juan Pino, Anne Wu and Jiatao Gu	4197
<i>A Visually-Grounded Parallel Corpus with Phrase-to-Region Linking</i>	
Hideki Nakayama, Akihiro Tamura and Takashi Ninomiya	4204
<i>Multilingual Dictionary Based Construction of Core Vocabulary</i>	
Winston Wu, Garrett Nicolai and David Yarowsky	4211
<i>Common Voice: A Massively-Multilingual Speech Corpus</i>	
Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers and Gregor Weber	4218
<i>Massively Multilingual Pronunciation Modeling with WikiPron</i>	
Jackson L. Lee, Lucas F.E. Ashby, M. Elizabeth Garza, Yeonju Lee-Sikka, Sean Miller, Alan Wong, Arya D. McCarthy and Kyle Gorman	4223
<i>HELFI: a Hebrew-Greek-Finnish Parallel Bible Corpus with Cross-Lingual Morpheme Alignment</i>	
Anssi Yli-Jyrä, Josi Purhonen, Matti Liljeqvist, Arto Antturi, Pekka Nieminen, Kari M. Rântilä and Valtter Luoto	4229
<i>ArzEn: A Speech Corpus for Code-switched Egyptian Arabic-English</i>	
Injy Hamed, Ngoc Thang Vu and Slim Abdennadher	4237
<i>Cross-lingual Named Entity List Search via Transliteration</i>	
Aleksandr Khakhmovich, Svetlana Pavlova, Kira Kirillova, Nikolay Arefyev and Ekaterina Savilova	4247

Multimodality

- Serial Speakers: a Dataset of TV Series*
Xavier Bost, Vincent Labatut and Georges Linares 4256
- Image Position Prediction in Multimodal Documents*
Masayasu Muraoka, Ryosuke Kohita and Etsuko Ishii 4265
- Visual Grounding Annotation of Recipe Flow Graph*
Taichi Nishimura, Suzushi Tomori, Hayato Hashimoto, Atsushi Hashimoto, Yoko Yamakata, Jun Harashima, Yoshitaka Ushiku and Shinsuke Mori 4275
- Building a Multimodal Entity Linking Dataset From Tweets*
Omar Adjali, Romaric Besançon, Olivier Ferret, Hervé Le Borgne and Brigitte Grau 4285
- A Multimodal Educational Corpus of Oral Courses: Annotation, Analysis and Case Study*
salima mdhaffar, Yannick Estève, Antoine Laurent, Nicolas Hernandez, Richard Dufour, Delphine Charlet, Geraldine Damnati, Solen Quiniou and Nathalie Camelin 4293
- Annotating Event Appearance for Japanese Chess Commentary Corpus*
Hirotaka Kameko and Shinsuke Mori 4302
- Offensive Video Detection: Dataset and Baseline Results*
Cleber Alcântara, Viviane Moreira and Diego Feijo 4309
- Adding Gesture, Posture and Facial Displays to the PoliModal Corpus of Political Interviews*
Daniela Trotta, Alessio Palmero Aprosio, Sara Tonelli and Annibale Elia 4320
- E:Calm Resource: a Resource for Studying Texts Produced by French Pupils and Students*
Lydia-Mai Ho-Dac, Serge Fleury and Claude Ponton 4327
- Introducing MULAI: A Multimodal Database of Laughter during Dyadic Interactions*
Michel-Pierre Jansen, Khiet P. Truong, Dirk K.J. Heylen and Deniece S. Nazareth 4333
- The Connection between the Text and Images of News Articles: New Insights for Multimedia Analysis*
Nelleke Oostdijk, Hans van Halteren, Erkan Başar and Martha Larson 4343
- LifeQA: A Real-life Dataset for Video Question Answering*
Santiago Castro, Mahmoud Azab, Jonathan Stroud, Cristina Noujaim, Ruoyao Wang, Jia Deng and Rada Mihalcea 4352

MultiWord Expressions & Collocations

- A Domain-Specific Dataset of Difficulty Ratings for German Noun Compounds in the Domains DIY, Cooking and Automotive*
Julia Bettinger, Anna Hättü, Michael Dorna and Sabine Schulte im Walde 4359
- All That Glitters is Not Gold: A Gold Standard of Adjective-Noun Collocations for German*
Yana Strakatova, Neele Falk, Isabel Fuhrmann, Erhard Hinrichs and Daniela Rossmann 4368
- Variants of Vector Space Reductions for Predicting the Compositionality of English Noun Compounds*
Pegah Alipoor and Sabine Schulte im Walde 4379

<i>Varying Vector Representations and Integrating Meaning Shifts into a PageRank Model for Automatic Term Extraction</i>	
Anurag Nigam, Anna Hätty and Sabine Schulte im Walde	4388
<i>Rigor Mortis: Annotating MWEs with a Gamified Platform</i>	
Karèn Fort, Bruno Guillaume, Yann-Alan Pilatte, Mathieu Constant and Nicolas Lefèbvre ...	4395
<i>A Multi-word Expression Dataset for Swedish</i>	
Murathan Kurfalı, Robert Östling, Johan Sjons and Mats Wirén	4402
<i>A Joint Approach to Compound Splitting and Idiomatic Compound Detection</i>	
Irina Krotova, Sergey Aksenov and Ekaterina Artemova	4410
<i>Dedicated Language Resources for Interdisciplinary Research on Multiword Expressions: Best Thing since Sliced Bread</i>	
Ferdy Hubers, Catia Cucchiarini and Helmer Strik	4418
<i>Detecting Multiword Expression Type Helps Lexical Complexity Assessment</i>	
Ekaterina Kochmar, Sian Gooding and Matthew Shardlow	4426
Named Entity Recognition	
<i>Introducing RONEC - the Romanian Named Entity Corpus</i>	
Stefan Daniel Dumitrescu and Andrei-Marius Avram	4436
<i>A Semi-supervised Approach for De-identification of Swedish Clinical Text</i>	
Hanna Berg and Hercules Dalianis	4444
<i>A Chinese Corpus for Fine-grained Entity Typing</i>	
Chin Lee, Hongliang Dai, Yangqiu Song and Xin Li	4451
<i>Czech Historical Named Entity Corpus v 1.0</i>	
Helena Hubková, Pavel Kral and Eva Pettersson	4458
<i>Code Alltag 2.0 — A Pseudonymized German-Language Email Corpus</i>	
Elisabeth Eder, Ulrike Krieg-Holz and Udo Hahn	4466
<i>A Dataset of German Legal Documents for Named Entity Recognition</i>	
Elena Leitner, Georg Rehm and Julian Moreno-Schneider	4478
<i>Sensitive Data Detection and Classification in Spanish Clinical Text: Experiments with BERT</i>	
Aitor García Pablos, Naiara Perez and Montse Cuadros	4486
<i>Named Entities in Medical Case Reports: Corpus and Experiments</i>	
Sarah Schulz, Jurica Ševa, Samuel Rodriguez, Malte Ostendorff and Georg Rehm	4495
<i>Hedwig: A Named Entity Linker</i>	
Marcus Klang and Pierre Nugues	4501
<i>An Experiment in Annotating Animal Species Names from ISTEK Resources</i>	
Sabine Barreaux and Dominique Besagni	4509
<i>Where are we in Named Entity Recognition from Speech?</i>	
Antoine Caubrière, Sophie Rosset, Yannick Estève, Antoine Laurent and Emmanuel Morin ..	4514

<i>Tagging Location Phrases in Text</i>	
Paul McNamee, James Mayfield, Cash Costello, Caitlyn Bishop and Shelby Anderson	4521
<i>ScienceExamCER: A High-Density Fine-Grained Science-Domain Corpus for Common Entity Recognition</i>	
Hannah Smith, Zeyu Zhang, John Culnan and Peter Jansen	4529
<i>NorNE: Annotating Named Entities for Norwegian</i>	
Fredrik Jørgensen, Tobias Aasmoe, Anne-Stine Ruud Husevåg, Lilja Øvrelid and Erik Velldal	4547
<i>Tag Me If You Can! Semantic Annotation of Biodiversity Metadata with the QEMP Corpus and the BiodivTagger</i>	
Felicitas Löffler, Nora Abdelmageed, Samira Babalou, Pawandeep Kaur and Birgitta König-Ries	4557
<i>Towards a Versatile Medical-Annotation Guideline Feasible Without Heavy Medical Knowledge: Starting From Critical Lung Diseases</i>	
Shuntaro Yada, Ayami Joh, Ribeka Tanaka, Fei Cheng, Eiji Aramaki and Sadao Kurohashi . .	4565
<i>Creating a Dataset for Named Entity Recognition in the Archaeology Domain</i>	
Alex Brandsen, Suzan Verberne, Milco Wansleebe and Karsten Lambers	4573
<i>Development of a Medical Incident Report Corpus with Intention and Factuality Annotation</i>	
Hongkuan Zhang, Ryohei Sasano, Koichi Takeda and Zoie Shui-Yee Wong	4578
<i>ProGene - A Large-scale, High-Quality Protein-Gene Annotated Benchmark Corpus</i>	
Erik Faessler, Luise Modersohn, Christina Lohr and Udo Hahn	4585
<i>DaNE: A Named Entity Resource for Danish</i>	
Rasmus Hvingelby, Amalie Brogaard Pauli, Maria Barrett, Christina Rosted, Lasse Malm Lidegaard and Anders Sjøgaard	4597
<i>Fine-grained Named Entity Annotations for German Biographic Interviews</i>	
Josef Ruppenhofer, Ines Rehbein and Carolina Flinz	4605
<i>A Broad-coverage Corpus for Finnish Named Entity Recognition</i>	
Jouni Luoma, Miika Oinonen, Maria Pyykönen, Veronika Laippala and Sampo Pyysalo	4615
<i>Embeddings for Named Entity Recognition in Geoscience Portuguese Literature</i>	
Bernardo Consoli, Joaquim Santos, Diogo Gomes, Fabio Cordeiro, Renata Vieira and Viviane Moreira	4625
<i>Establishing a New State-of-the-Art for French Named Entity Recognition</i>	
Pedro Javier Ortiz Suárez, Yoann Dupont, Benjamin Muller, Laurent Romary and Benoît Sagot	4631
<i>Building OCR/NER Test Collections</i>	
Dawn Lawrie, James Mayfield and David Etter	4639
<i>Reconstructing NER Corpora: a Case Study on Bulgarian</i>	
Iva Marinova, Laska Laskova, Petya Osenova, Kiril Simov and Alexander Popov	4647

Natural Language Generation

<i>MucLex: A German Lexicon for Surface Realisation</i> Kira Klimt, Daniel Braun, Daniela Schneider and Florian Matthes	4653
<i>Generating Major Types of Chinese Classical Poetry in a Uniformed Framework</i> Jinyi Hu and Maosong Sun	4658
<i>Video Caption Dataset for Describing Human Actions in Japanese</i> Yutaro Shigeto, Yuya Yoshikawa, Jiaqing Lin and Akikazu Takeuchi	4664
<i>Decode with Template: Content Preserving Sentiment Transfer</i> Zhiyuan Wen, Jiannong Cao, Ruosong Yang and Senzhang Wang	4671
<i>Best Student Forcing: A Simple Training Mechanism in Adversarial Language Generation</i> Jonathan Sauder, Ting Hu, Xiaoyin Che, Goncalo Mordido, Haojin Yang and Christoph Meinel	4680
<i>Controllable Sentence Simplification</i> Louis Martin, Éric de la Clergerie, Benoît Sagot and Antoine Bordes	4689
<i>Exploring Transformer Text Generation for Medical Dataset Augmentation</i> Ali Amin-Nejad, Julia Ive and Sumithra Velupillai	4699
<i>Multi-lingual Mathematical Word Problem Generation using Long Short Term Memory Networks with Enhanced Input Features</i> Vijini Liyanage and Surangika Ranathunga	4709

Neural language representation models

<i>Time-Aware Word Embeddings for Three Lebanese News Archives</i> Jad Doughman, Fatima Abu Salem and Shady Elbassuoni	4717
<i>GGP: Glossary Guided Post-processing for Word Embedding Learning</i> Ruosong Yang, Jiannong Cao and Zhiyuan Wen	4726
<i>High Quality ELMo Embeddings for Seven Less-Resourced Languages</i> Matej Ulčar and Marko Robnik-Šikonja	4731
<i>Is Language Modeling Enough? Evaluating Effective Embedding Combinations</i> Rudolf Schneider, Tom Oberhauser, Paul Grundmann, Felix Alexander Gers, Alexander Loeser and Steffen Staab	4739
<i>Language Modeling with a General Second-Order RNN</i> Diego Maupomé and Marie-Jean Meurs	4749
<i>Towards a Gold Standard for Evaluating Danish Word Embeddings</i> Nina Schneidermann, Rasmus Hvingelby and Bolette Pedersen	4754
<i>Urban Dictionary Embeddings for Slang NLP Applications</i> Steven Wilson, Walid Magdy, Barbara McGillivray, Kiran Garimella and Gareth Tyson	4764

<i>Representation Learning for Unseen Words by Bridging Subwords to Semantic Networks</i> Yeachan Kim, Kang-Min Kim and SangKeun Lee	4774
<i>Give your Text Representation Models some Love: the Case for Basque</i> Rodrigo Agerri, Iñaki San Vicente, Jon Ander Campos, Ander Barrena, Xabier Saralegi, Aitor Soroa and Eneko Agirre	4781
<i>On the Correlation of Word Embedding Evaluation Metrics</i> François Torregrossa, Vincent Claveau, Nihel Kooli, Guillaume Gravier and Robin Allesiardo	4789
<i>CBOw-tag: a Modified CBOw Algorithm for Generating Embedding Models from Annotated Corpora</i> Attila Novák, László Laki and Borbála Novák	4798
<i>Much Ado About Nothing – Identification of Zero Copulas in Hungarian Using an NMT Model</i> Andrea Dömötör, Zijian Győző Yang and Attila Novák	4802
<i>Leveraging Contextual Embeddings for Detecting Diachronic Semantic Shift</i> Matej Martinc, Petra Kralj Novak and Senja Pollak	4811
<i>Improving NMT Quality Using Terminology Injection</i> Duane K. Dougal and Deryle Lonsdale	4820
<i>Word Embedding Evaluation in Downstream Tasks and Semantic Analogies</i> Joaquim Santos, Bernardo Consoli and Renata Vieira	4828
<i>Detection of Reading Absorption in User-Generated Book Reviews: Resources Creation and Evaluation</i> Piroska Lendvai, Sándor Darányi, Christian Geng, Moniek Kuijpers, Oier Lopez de Lacalle, Jean-Christophe Mensonides, Simone Reborá and Uwe Reichel	4835
Ontologies and Wordnet	
<i>Developing an Arabic Infectious Disease Ontology to Include Non-Standard Terminology</i> Lama Alsudias and Paul Rayson	4842
<i>Aligning Wikipedia with WordNet: a Review and Evaluation of Different Techniques</i> Antoni Oliver	4851
<i>The MWN.PT WordNet for Portuguese: Projection, Validation, Cross-lingual Alignment and Distribution</i> António Branco, Sara Grilo, Márcia Bolrinha, Chakaveh Saedi, Ruben Branco, João Silva, Andreia Querido, Rita de Carvalho, Rosa Gaudio, Mariana Avelãs and Clara Pinto	4859
<i>Ontology-Style Relation Annotation: A Case Study</i> Savong Bou, Naoki Suzuki, Makoto Miwa and Yutaka Sasaki	4867
<i>The Ontology of Bulgarian Dialects – Architecture and Information Retrieval</i> Rositsa Dekova	4877
<i>Spatial AMR: Expanded Spatial Annotation in the Context of a Grounded Minecraft Corpus</i> Julia Bonn, Martha Palmer, Zheng Cai and Kristin Wright-Bettner	4883
<i>English WordNet Random Walk Pseudo-Corpora</i> Filip Klubička, Alfredo Maldonado, Abhijit Mahalunkar and John Kelleher	4893

<i>On the Formal Standardization of Terminology Resources: The Case Study of TriMED</i>	
Federica Vezzani and Giorgio Maria Di Nunzio	4903

Opinion Mining , Sentiment Analysis

<i>Metaphorical Expressions in Automatic Arabic Sentiment Analysis</i>	
Israa Alsiyat and Scott Piao	4911
<i>HotelRec: a Novel Very Large-Scale Hotel Recommendation Dataset</i>	
Diego Antognini and Boi Faltings	4917
<i>Doctor Who? Framing Through Names and Titles in German</i>	
Esther van den Berg, Katharina Korfhage, Josef Ruppenhofer, Michael Wiegand and Katja Markert	4924
<i>Adapt or Get Left Behind: Domain Adaptation through BERT Language Model Finetuning for Aspect-Target Sentiment Classification</i>	
Alexander Rietzler, Sebastian Stabinger, Paul Opitz and Stefan Engl	4933
<i>An Empirical Examination of Online Restaurant Reviews</i>	
Hyun Jung Kang and Iris Eshkol-Taravella	4942
<i>Manovaad: A Novel Approach to Event Oriented Corpus Creation Capturing Subjectivity and Focus</i>	
Lalitha Kameswari and Radhika Mamidi	4948
<i>Toward Qualitative Evaluation of Embeddings for Arabic Sentiment Analysis</i>	
Amira Barhoumi, Nathalie Camelin, Chafik Aloulou, Yannick Estève and Lamia Hadrich Belguith	4955
<i>Annotating Perspectives on Vaccination</i>	
Roser Morante, Chantal van Son, Isa Maks and Piek Vossen	4964
<i>Aspect On: an Interactive Solution for Post-Editing the Aspect Extraction based on Online Learning</i>	
Mara China-Rios, Marc Franco-Salvador and Yassine Benajiba	4974
<i>Recommendation Chart of Domains for Cross-Domain Sentiment Analysis: Findings of A 20 Domain Study</i>	
Akash Sheoran, Diptesh Kanojia, Aditya Joshi and Pushpak Bhattacharyya	4982
<i>Inference Annotation of a Chinese Corpus for Opinion Mining</i>	
Liyun Yan, Danni E, Mei Gan, Cyril Grouin and Mathieu Valette	4991
<i>Cooking Up a Neural-based Model for Recipe Classification</i>	
Elham Mohammadi, Nada Naji, Louis Marceau, Marc Queudot, Eric Charton, Leila Kosseim and Marie-Jean Meurs	5000
<i>Enhancing a Lexicon of Polarity Shifters through the Supervised Classification of Shifting Directions</i>	
Marc Schulder, Michael Wiegand and Josef Ruppenhofer	5010
<i>Dataset Creation and Evaluation of Aspect Based Sentiment Analysis in Telugu, a Low Resource Language</i>	
Yashwanth Reddy Regatte, Rama Rohit Reddy Gangula and Radhika Mamidi	5017

<i>A Fine-grained Sentiment Dataset for Norwegian</i>	
Lilja Øvrelid, Petter Mæhlum, Jeremy Barnes and Erik Velldal	5025
<i>The Design and Construction of a Chinese Sarcasm Dataset</i>	
Xiaochang Gong, Qin Zhao, Jun Zhang, Ruibin Mao and Ruifeng Xu	5034
<i>Target-based Sentiment Annotation in Chinese Financial News</i>	
Chaofa Yuan, Yuhan Liu, Rongdi Yin, Jun Zhang, Qinling Zhu, Ruibin Mao and Ruifeng Xu	5040
<i>Multi-domain Tweet Corpora for Sentiment Analysis: Resource Creation and Evaluation</i>	
Mamta ., Asif Ekbal, Pushpak Bhattacharyya, Shikha Srivastava, Alka Kumar and Tista Saha	5046
<i>Reproduction and Revival of the Argument Reasoning Comprehension Task</i>	
João António Rodrigues, Ruben Branco, João Silva and António Branco	5055
<i>Design and Evaluation of SentiEcon: a fine-grained Economic/Financial Sentiment Lexicon from a Corpus of Business News</i>	
Antonio Moreno-Ortiz, Javier Fernandez-Cruz and Chantal Pérez Chantal Hernández	5065
<i>ParlVote: A Corpus for Sentiment Analysis of Political Debates</i>	
Gavin Abercrombie and Riza Batista-Navarro	5073
<i>Offensive Language Detection Using Brown Clustering</i>	
Zuoyu Tian and Sandra Kübler	5079
<i>Annotating for Hate Speech: The MaNeCo Corpus and Some Input from Critical Discourse Analysis</i>	
Stavros Assimakopoulos, Rebecca Vella Muskat, Lonneke van der Plas and Albert Gatt	5088
<i>Marking Irony Activators in a Universal Dependencies Treebank: The Case of an Italian Twitter Corpus</i>	
Alessandra Teresa Cignarella, Manuela Sanguinetti, Cristina Bosco and Paolo Rosso	5098
<i>HAHA 2019 Dataset: A Corpus for Humor Analysis in Spanish</i>	
Luis Chiruzzo, Santiago Castro and Aiala Rosá	5106
<i>Offensive Language Identification in Greek</i>	
Zesis Pitenis, Marcos Zampieri and Tharindu Ranasinghe	5113
Parsing, Grammar, Syntax, Treebank	
<i>Syntax and Semantics in a Treebank for Esperanto</i>	
Eckhard Bick	5120
<i>Implementation and Evaluation of an LFG-based Parser for Wolof</i>	
Cheikh M. Bamba Dione	5128
<i>The Treebank of Vedic Sanskrit</i>	
Oliver Hellwig, Salvatore Scarlata, Elia Ackermann and Paul Widmer	5137
<i>Inherent Dependency Displacement Bias of Transition-Based Algorithms</i>	
Mark Anderson and Carlos Gómez-Rodríguez	5147
<i>A Gold Standard Dependency Treebank for Turkish</i>	
Tolga Kayadelen, Adnan Ozturel and Bernd Bohnet	5156

<i>Chunk Different Kind of Spoken Discourse: Challenges for Machine Learning</i> Iris Eshkol-Taravella, Mariame Maarouf, Flora Badin, Marie Skrovec and Isabelle Tellier	5164
<i>GRAIN-S: Manually Annotated Syntax for German Interviews</i> Agnieszka Falenska, Zoltán Czesznak, Kerstin Jung, Moritz Völkel, Wolfgang Seeker and Jonas Kuhn	5169
<i>Yorùbá Dependency Treebank (YTB)</i> Olájídé Ishola and Daniel Zeman	5178
<i>English Recipe Flow Graph Corpus</i> Yoko Yamakata, Shinsuke Mori and John Carroll	5187
<i>Development of a General-Purpose Categorical Grammar Treebank</i> Yusuke Kubota, Koji Mineshima, Noritsugu Hayashi and Shinya Okano	5195
<i>Dependency Parsing for Urdu: Resources, Conversions and Learning</i> Toqeer Ehsan and Miriam Butt	5202
<i>Prague Dependency Treebank - Consolidated 1.0</i> Jan Hajic, Eduard Bejček, Jaroslava Hlavacova, Marie Mikulová, Milan Straka, Jan Štěpánek and Barbora Štěpánková	5208
<i>Training a Swedish Constituency Parser on Six Incompatible Treebanks</i> Richard Johansson and Yvonne Adesam	5219
<i>Parsing as Tagging</i> Robert Vacareanu, George Caique Gouveia Barbosa, Marco A. Valenzuela-Escárcega and Mihai Surdeanu	5225
<i>The EDGeS Diachronic Bible Corpus</i> Gerlof Bouma, Evie Coussé, Trude Dijkstra and Nicoline van der Sijs	5232
<i>Treebanking User-Generated Content: A Proposal for a Unified Representation in Universal Dependencies</i> Manuela Sanguinetti, Cristina Bosco, Lauren Cassidy, Özlem Çetinoğlu, Alessandra Teresa Cignarella, Teresa Lynn, Ines Rehbein, Josef Ruppenhofer, Djamel Seddah and Amir Zeldes	5240
<i>A Diachronic Treebank of Russian Spanning More Than a Thousand Years</i> Aleksandrs Berdicevskis and Hanne Eckhoff	5251
<i>ÆTHEL: Automatically Extracted Typological Derivations for Dutch</i> Konstantinos Kogkalidis, Michael Moortgat and Richard Moot	5257
<i>GUMBY – A Free, Balanced, and Rich English Web Corpus</i> Luke Gessler, Siyao Peng, Yang Liu, Yilun Zhu, Shabnam Behzad and Amir Zeldes	5267
<i>Typical Sentences as a Resource for Valence</i> Uwe Quasthoff, Lars Hellan, Erik Körner, Thomas Eckart, Dirk Goldhahn and Dorothee Beermann	5276
<i>Recognizing Sentence-level Logical Document Structures with the Help of Context-free Grammars</i> Jonathan Hildebrand, Wahed Hemati and Alexander Mehler	5282
<i>When Collaborative Treebank Curation Meets Graph Grammars</i> Gaël Guibon, Marine Courtin, Kim Gerdes and Bruno Guillaume	5291

ODIL_Syntax: a Free Spontaneous Spoken French Treebank Annotated with Constituent Trees
Ilaine Wang, Aurore Pelletier, Jean-Yves Antoine and Anaïs Halftermeyer 5301

Towards the Conversion of National Corpus of Polish to Universal Dependencies
Alina Wróblewska 5308

Phonetic Databases, Phonology

SegBo: A Database of Borrowed Sounds in the World's Language
Eitan Grossman, Elad Eisen, Dmitry Nikolaev and Steven Moran 5316

Developing Resources for Automated Speech Processing of Quebec French
Mélanie Lancien, Marie-Hélène Côté and Brigitte Bigi 5323

AlloVera: A Multilingual Allophone Database
David R. Mortensen, Xinjian Li, Patrick Littell, Alexis Michaud, Shruti Rijhwani, Antonios Anastopoulos, Alan W Black, Florian Metze and Graham Neubig 5329

Arabic Speech Rhythm Corpus: Read and Spontaneous Speaking Styles
Omnia Ibrahim, Homa Asadi, Eman Kassem and Volker Dellwo 5337

Comparing Methods for Measuring Dialect Similarity in Norwegian
Janne Johannessen, Andre Kåsen, Kristin Hagen, Anders Nøklestad and Joel Priestley 5343

AccentDB: A Database of Non-Native English Accents to Assist Neural Speech Recognition
Afroz Ahamad, Ankit Anand and Pranesh Bhargava 5351

Question Answering

A Framework for Evaluation of Machine Reading Comprehension Gold Standards
Viktor Schlegel, Marco Valentino, Andre Freitas, Goran Nenadic and Riza Batista-Navarro .. 5359

Multi-class Hierarchical Question Classification for Multiple Choice Science Exams
Dongfang Xu, Peter Jansen, Jaycie Martin, Zhengnan Xie, Vikas Yadav, Harish Tayyar Madabushi, Oyvind Tafjord and Peter Clark 5370

Assessing Users' Reputation from Syntactic and Semantic Information in Community Question Answering
Yonas Woldemariam 5383

Unsupervised Domain Adaptation of Language Models for Reading Comprehension
Kosuke Nishida, Kyosuke Nishida, Itsumi Saito, Hisako Asano and Junji Tomita 5392

Propagate-Selector: Detecting Supporting Sentences for Question Answering via Graph Neural Networks
Seunghyun Yoon, Franck Dernoncourt, Doo Soon Kim, Trung Bui and Kyomin Jung 5400

An Empirical Comparison of Question Classification Methods for Question Answering Systems
Eduardo Cortes, Vinicius Woloszyn, Arne Binder, Tilo Himmelsbach, Dante Barone and Sebastian Möller 5408

<i>Cross-sentence Pre-trained Model for Interactive QA matching</i> Jinmeng Wu and Yanbin Hao	5417
<i>SQuAD2-CR: Semi-supervised Annotation for Cause and Rationales for Unanswerability in SQuAD 2.0</i> Gyeongbok Lee, Seung-won Hwang and Hyunsouk Cho	5425
<i>Generating Responses that Reflect Meta Information in User-Generated Question Answer Pairs</i> Takashi Kodama, Ryuichiro Higashinaka, Koh Mitsuda, Ryo Masumura, Yushi Aono, Ryuta Nakamura, Noritake Adachi and Hidetoshi Kawabata	5433
<i>AIA-BDE: A Corpus of FAQs in Portuguese and their Variations</i> Hugo Gonçalo Oliveira, João Ferreira, José Santos, Pedro Fialho, Ricardo Rodrigues, Luisa Coheur and Ana Alves	5442
<i>TutorialVQA: Question Answering Dataset for Tutorial Videos</i> Anthony Colas, Seokhwan Kim, Franck Dernoncourt, Siddhesh Gupte, Zhe Wang and Doo Soon Kim	5450
<i>WorldTree V2: A Corpus of Science-Domain Structured Explanations and Inference Patterns supporting Multi-Hop Inference</i> Zhengnan Xie, Sebastian Thiem, Jaycie Martin, Elizabeth Wainwright, Steven Marmorstein and Peter Jansen	5456
<i>Chat or Learn: a Data-Driven Robust Question-Answering System</i> Gabriel Luthier and Andrei Popescu-Belis	5474
<i>Project PIAF: Building a Native French Question-Answering Dataset</i> Rachel Keraron, Guillaume Lancrenon, Mathilde Bras, Frédéric Allary, Gilles Moyse, Thomas Sialom, Edmundo-Pavel Soriano-Morales and Jacopo Staiano	5481
<i>Cross-lingual and Cross-domain Evaluation of Machine Reading Comprehension with Squad and CALOR-Quest Corpora</i> Delphine Charlet, Geraldine Damnati, Frederic Bechet, gabriel marzinotto and Johannes Heinecke	5491
<i>ScholarlyRead: A New Dataset for Scientific Article Reading Comprehension</i> Tanik Saikh, Asif Ekbal and Pushpak Bhattacharyya	5498
<i>Contextualized Embeddings based Transformer Encoder for Sentence Similarity Modeling in Answer Selection Task</i> Md Tahmid Rahman Laskar, Jimmy Xiangji Huang and Enamul Hoque	5505
<i>Automatic Spanish Translation of SQuAD Dataset for Multi-lingual Question Answering</i> Casimiro Pio Carrino, Marta R. Costa-jussà and José A. R. Fonollosa	5515
<i>A Corpus for Visual Question Answering Annotated with Frame Semantic Information</i> Mehrdad Alizadeh and Barbara Di Eugenio	5524
<i>Evaluation of Dataset Selection for Pre-Training and Fine-Tuning Transformer Language Models for Clinical Question Answering</i> Sarvesh Soni and Kirk Roberts	5532

REPROLANG2020 Track

A Shared Task of a New, Collaborative Type to Foster Reproducibility: A First Exercise in the Area of Language Science and Technology with REPROLANG2020

António Branco, Nicoletta Calzolari, Piek Vossen, Gertjan Van Noord, Dieter van Uytvanck, João Silva, Luís Gomes, André Moreira and Willem Elbers 5539

A Robust Self-Learning Method for Fully Unsupervised Cross-Lingual Mappings of Word Embeddings: Making the Method Robustly Reproducible as Well

Nicolas Garneau, Mathieu Godbout, David Beauchemin, Audrey Durand and Luc Lamontagne 5546

A Closer Look on Unsupervised Cross-lingual Word Embeddings Mapping

Kamil Pluciński, Mateusz Lango and Michał Zimmiewicz 5555

Reproducing a Morphosyntactic Tagger with a Meta-BiLSTM Model over Context Sensitive Token Encodings

Yung Han Khoe 5563

Reproducing Neural Ensemble Classifier for Semantic Relation Extraction in Scientific Papers

Kyeongmin Rim, Jingxuan Tu, Kelley Lynch and James Pustejovsky 5569

ULMFiT replication

Mohamed Abdellatif and Ahmed Elgammal 5579

CombiNMT: An Exploration into Neural Text Simplification Models

Michael Cooper and Matthew Shardlow 5588

Reproducing Monolingual, Multilingual and Cross-Lingual CEFR Predictions

Yves Bestgen 5595

Reproduction and Replication: A Case Study with Automatic Essay Scoring

Eva Huber and Çağrı Çöltekin 5603

REPROLANG 2020: Automatic Proficiency Scoring of Czech, English, German, Italian, and Spanish Learner Essays

Andrew Caines and Paula Buttery 5614

Language Proficiency Scoring

Cristina Arhiliuc, Jelena Mitrović and Michael Granitzer 5624

The Learnability of the Annotated Input in NMT Replicating (Vanmassenhove and Way, 2018) with Open-NMT

Nicolas Ballier, Nabil Amari, Laure Merat and Jean-Baptiste Yunès 5631

Semantic Web and Linked Data

KGvec2go – Knowledge Graph Embeddings as a Service

Jan Portisch, Michael Hladik and Heiko Paulheim 5641

Ontology Matching Using Convolutional Neural Networks

Alexandre Bento, Amal Zouaq and Michel Gagnon 5648

<i>Defying Wikidata: Validation of Terminological Relations in the Web of Data</i>	
Patricia Martín-Chozas, Sina Ahmadi and Elena Montiel-Ponsoda	5654
<i>Recent Developments for the Linguistic Linked Open Data Infrastructure</i>	
Thierry Declerck, John Philip McCrae, Matthias Hartung, Jorge Gracia, Christian Chiarcos, Elena Montiel-Ponsoda, Philipp Cimiano, Artem Revenko, Roser Saurí, Deirdre Lee, Stefania Racioppa, Jamal Abdul Nasir, Matthias Orlikowsk, Marta Lanau-Coronas, Christian Fäth, Mariano Rico, Mohammad Fazleh Elahi, Maria Khvalchik, Meritxell Gonzalez and Katharine Cooney	5660
<i>Annotation Interoperability for the Post-ISOCat Era</i>	
Christian Chiarcos, Christian Fäth and Frank Abromeit	5668
Semantics	
<i>A Large Harvested Corpus of Location Metonymy</i>	
Kevin Alex Mathews and Michael Strube	5678
<i>The DAPRECO Knowledge Base: Representing the GDPR in LegalRuleML</i>	
Livio Robaldo, Cesare Bartolini and Gabriele Lenzini	5688
<i>The Universal Decompositional Semantics Dataset and Decomp Toolkit</i>	
Aaron Steven White, Elias Stengel-Eskin, Siddharth Vashishtha, Venkata Subrahmanyam Govindarajan, Dee Ann Reisinger, Tim Vieira, Keisuke Sakaguchi, Sheng Zhang, Francis Ferraro, Rachel Rudinger, Kyle Rawlins and Benjamin Van Durme	5698
<i>Are Word Embeddings Really a Bad Fit for the Estimation of Thematic Fit?</i>	
Emmanuele Chersoni, Ludovica Pannitto, Enrico Santus, Alessandro Lenci and Chu-Ren Huang	5708
<i>Ciron: a New Benchmark Dataset for Chinese Irony Detection</i>	
Rong Xiang, Xuefeng Gao, Yunfei Long, Anran Li, Emmanuele Chersoni, Qin Lu and Chu-Ren Huang	5714
<i>wikiHowToImprove: A Resource and Analyses on Edits in Instructional Texts</i>	
Talita Anthonio, Irshad Bhat and Michael Roth	5721
<i>Must Children be Vaccinated or not? Annotating Modal Verbs in the Vaccination Debate</i>	
Liza King and Roser Morante	5730
<i>NegBERT: A Transfer Learning Approach for Negation Detection and Scope Resolution</i>	
Aditya Khandelwal and Suraj Sawant	5739
<i>Spatial Multi-Arrangement for Clustering and Multi-way Similarity Dataset Construction</i>	
Olga Majewska, Diana McCarthy, Jasper van den Bosch, Nikolaus Kriegeskorte, Ivan Vulić and Anna Korhonen	5749
<i>A Short Survey on Sense-Annotated Corpora</i>	
Tommaso Pasini and Jose Camacho-Collados	5759
<i>Using Distributional Thesaurus Embedding for Co-hyponymy Detection</i>	
Abhik Jana, Nikhil Reddy Varimalla and Pawan Goyal	5766

<i>NUBes: A Corpus of Negation and Uncertainty in Spanish Clinical Texts</i> Salvador Lima Lopez, Naiara Perez, Montse Cuadros and German Rigau	5772
<i>Decomposing and Comparing Meaning Relations: Paraphrasing, Textual Entailment, Contradiction, and Specificity</i> Venelin Kovatchev, Darina Gold, M. Antonia Marti, Maria Salamo and Torsten Zesch	5782
<i>Object Naming in Language and Vision: A Survey and a New Dataset</i> Carina Silberer, Sina Zarri� and Gemma Boleda	5792
<i>MSD-1030: A Well-built Multi-Sense Evaluation Dataset for Sense Representation Models</i> Ting-Yu Yen, Yang-Yin Lee, Yow-Ting Shiue, Hen-Hsen Huang and Hsin-Hsi Chen	5802
<i>Figure Me Out: A Gold Standard Dataset for Metaphor Interpretation</i> Omnia Zayed, John Philip McCrae and Paul Buitelaar	5810
<i>Extrinsic Evaluation of French Dependency Parsers on a Specialized Corpus: Comparison of Distributional Thesauri</i> Ludovic Tanguy, Pauline Brunet and Olivier Ferret	5820
<i>Dataset and Enhanced Model for Eligibility Criteria-to-SQL Semantic Parsing</i> Xiaojing Yu, Tianlong Chen, Zhengjie Yu, Huiyu Li, Yang Yang, Xiaoqian Jiang and Anxiao Jiang	5829
<i>Recognizing Semantic Relations by Combining Transformers and Fully Connected Models</i> Dmitri Roussinov, Serge Sharoff and Nadezhda Puchmina	5838
<i>Word Attribute Prediction Enhanced by Lexical Entailment Tasks</i> Mika Hasegawa, Tetsunori Kobayashi and Yoshihiko Hayashi	5846
<i>From Spatial Relations to Spatial Configurations</i> Soham Dan, Parisa Kordjamshidi, Julia Bonn, Archana Bhatia, Zheng Cai, Martha Palmer and Dan Roth	5855
<i>Representing Verbs with Visual Argument Vectors</i> Irene Sucameli and Alessandro Lenci	5865
<i>Are White Ravens Ever White? - Non-Literal Adjective-Noun Phrases in Polish</i> Agnieszka Mykowiecka and Malgorzata Marciniak	5871
<i>CoSimLex: A Resource for Evaluating Graded Word Similarity in Context</i> Carlos Santos Armendariz, Matthew Purver, Matej Ul�ar, Senja Pollak, Nikola Ljube�i� and Mark Granroth-Wilding	5878
<i>A French Version of the FraCaS Test Suite</i> Maxime Amblard, Cl�ment Beysson, Philippe de Groote, Bruno Guillaume and Sylvain Pogodalla	5887
<i>Automatic Compilation of Resources for Academic Writing and Evaluating with Informal Word Identifi- cation and Paraphrasing System</i> Seid Muhie Yimam, Gopalakrishnan Venkatesh, John Lee and Chris Biemann	5896
<i>Sense-Annotated Corpora for Word Sense Disambiguation in Multiple Languages and Domains</i> Bianca Scarlini, Tommaso Pasini and Roberto Navigli	5905

<i>FrSemCor: Annotating a French Corpus with Supersenses</i> Lucie Barque, Pauline Haas, Richard Huyghe, Delphine Tribout, Marie Candito, Benoit Crabbé and Vincent Segonne	5912
<i>A Formal Analysis of Multimodal Referring Strategies Under Common Ground</i> Nikhil Krishnaswamy and James Pustejovsky	5919
<i>Improving Neural Metaphor Detection with Visual Datasets</i> Gitit Kehat and James Pustejovsky	5928
<i>Building a Hebrew Semantic Role Labeling Lexical Resource from Parallel Movie Subtitles</i> Ben Eyal and Michael Elhadad	5934
<i>Word Sense Disambiguation for 158 Languages using Word Embeddings Only</i> Varvara Logacheva, Denis Teslenko, Artem Shelmanov, Steffen Remus, Dmitry Ustalov, Andrey Kutuzov, Ekaterina Artemova, Chris Biemann, Simone Paolo Ponzetto and Alexander Panchenko ..	5943
<i>Extraction of Hyponymic Relations in French with Knowledge-Pattern-Based Word Sketches</i> Antonio San Martín, Catherine Trekker and Pilar León-Araúz	5953
<i>SeCoDa: Sense Complexity Dataset</i> David Strohmaier, Sian Gooding, Shiva Taslimipoor and Ekaterina Kochmar	5962
<i>A New Resource for German Causal Language</i> Ines Rehbein and Josef Ruppenhofer	5968
<i>One Classifier for All Ambiguous Words: Overcoming Data Sparsity by Utilizing Sense Correlations Across Words</i> Prafulla Kumar Choubey and Ruihong Huang	5978
<i>A Corpus of Adpositional Supersenses for Mandarin Chinese</i> Siyao Peng, Yang Liu, Yilun Zhu, Austin Blodgett, Yushi Zhao and Nathan Schneider	5986
<i>The Russian PropBank</i> Sarah Moeller, Irina Wagner, Martha Palmer, Kathryn Conger and Skatje Myers	5995
Sign Language Recognition and Generation	
<i>What Comes First: Combining Motion Capture and Eye Tracking Data to Study the Order of Articulators in Constructed Action in Sign Language Narratives</i> Tommi Jantunen, Anna Puupponen and Birgitta Burger	6003
<i>LSF-ANIMAL: A Motion Capture Corpus in French Sign Language Designed for the Animation of Signing Avatars</i> Lucie Naert, Caroline Larboulette and Sylvie Gibet	6008
<i>Sign Language Recognition with Transformer Networks</i> Mathieu De Coster, Mieke Van Herreweghe and Joni Dambre	6018
<i>Annotating a Fable in Italian Sign Language (LIS)</i> Serena Trolvi and Rodolfo Delmonte	6025
<i>HamNoSys2SiGML: Translating HamNoSys Into SiGML</i> Carolina Neves, Luísa Coheur and Hugo Nicolau	6035

<i>Dicta-Sign-LSF-v2: Remake of a Continuous French Sign Language Dialogue Corpus and a First Baseline for Automatic Sign Language Processing</i>	Valentin Belissen, Annelies Braffort and Michèle Gouiffès	6040
<i>An HMM Approach with Inherent Model Selection for Sign Language and Gesture Recognition</i>	Sandrine Tornay, Oya Aran and Mathew Magimai Doss	6049
<i>VROAV: Using Iconicity to Visually Represent Abstract Verbs</i>	Simone Scicluna and Carlo Strapparava	6057
<i>MEDI-API-SKEL - A 2D-Skeleton Video Database of French Sign Language With Aligned French Subtitles</i>	Hannah Bull, Annelies Braffort and Michèle Gouiffès	6063
<i>Alignment Data base for a Sign Language Concordancer</i>	Marion Kaczmarek and Michael Filhol	6069
<i>Evaluation of Manual and Non-manual Components for Sign Language Recognition</i>	Medet Mukushev, Arman Sabyrov, Alfarabi Imashev, Kenessary Koishybay, Vadim Kimmelman and Anara Sandygulova	6073
<i>TheRuSLan: Database of Russian Sign Language</i>	Ildar Kagirov, Denis Ivanko, Dmitry Ryumin, Alexander Axyonov and Alexey Karpov	6079
Social Media Processing		
<i>A Survey on Natural Language Processing for Fake News Detection</i>	Ray Oshikawa, Jing Qian and William Yang Wang	6086
<i>RP-DNN: A Tweet Level Propagation Context Based Deep Neural Networks for Early Rumor Detection in Social Media</i>	Jie Gao, Sooji Han, Xingyi Song and Fabio Ciravegna	6094
<i>Issues and Perspectives from 10,000 Annotated Financial Social Media Data</i>	Chung-Chi Chen, Hen-Hsen Huang and Hsin-Hsi Chen	6106
<i>Searching Brazilian Twitter for Signs of Mental Health Issues</i>	Wesley Santos, Amanda Funabashi and Ivandré Paraboni	6111
<i>RedDust: a Large Reusable Dataset of Reddit User Traits</i>	Anna Tigunova, Paramita Mirza, Andrew Yates and Gerhard Weikum	6118
<i>An Annotated Social Media Corpus for German</i>	Eckhard Bick	6127
<i>The rJokes Dataset: a Large Scale Humor Collection</i>	Orion Weller and Kevin Seppi	6136
<i>EmpiriST Corpus 2.0: Adding Manual Normalization, Lemmatization and Semantic Tagging to a German Web and CMC Corpus</i>	Thomas Proisl, Natalie Dykes, Philipp Heinrich, Besim Kabashi, Andreas Blombach and Stefan Evert	6142

<i>Fakeddit: A New Multimodal Benchmark Dataset for Fine-grained Fake News Detection</i> Kai Nakamura, Sharon Levy and William Yang Wang	6149
<i>Optimising Twitter-based Political Election Prediction with Relevance and Sentiment Filters</i> Eric Sanders and Antal van den Bosch	6158
<i>A Real-Time System for Credibility on Twitter</i> Adrian Iftene, Daniela Gifu, Andrei-Remus Miron and Mihai-Stefan Dudu	6166
<i>A Corpus of Turkish Offensive Language on Social Media</i> Çağrı Çöltekin	6174
<i>From Witch's Shot to Music Making Bones - Resources for Medical Laymen to Technical Language and Vice Versa</i> Laura Seiffe, Oliver Marten, Michael Mikhailov, Sven Schmeier, Sebastian Möller and Roland Roller	6185
<i>I Feel Offended, Don't Be Abusive! Implicit/Explicit Messages in Offensive and Abusive Language</i> Tommaso Caselli, Valerio Basile, Jelena Mitrović, Inga Kartoziya and Michael Granitzer	6193
<i>A Multi-Platform Arabic News Comment Dataset for Offensive Language Detection</i> Shammur Absar Chowdhury, Hamdy Mubarak, Ahmed Abdelali, Soon-gyo Jung, Bernard J Jansen and Joni Salminen	6203
<i>Twitter Trend Extraction: A Graph-based Approach for Tweet and Hashtag Ranking, Utilizing No-Hashtag Tweets</i> Zahra Majdabadi, Behnam Sabeti, Preni Golazizian, Seyed Arad Ashrafi Asli, Omid Momenzadeh and reza fahmi	6213
<i>A French Corpus for Event Detection on Twitter</i> Béatrice Mazoyer, Julia Cagé, Nicolas Hervé and Céline Hudelot	6220
<i>Minority Positive Sampling for Switching Points - an Anecdote for the Code-Mixing Language Modeling</i> Arindam Chatterjere, Vineeth Gupta, Parul Chopra and Amitava Das	6228
<i>Do You Really Want to Hurt Me? Predicting Abusive Swearing in Social Media</i> Endang Wahyu Pamungkas, Valerio Basile and Viviana Patti	6237
<i>Detecting Troll Tweets in a Bilingual Corpus</i> Lin Miao, Mark Last and Marina Litvak	6247
<i>Collecting Tweets to Investigate Regional Variation in Canadian English</i> Filip Miletic, Anne Przewozny-Desrioux and Ludovic Tanguy	6255
<i>DAICT: A Dialectal Arabic Irony Corpus Extracted from Twitter</i> Ines Abbes, Wajdi Zaghouni, Omaira El-Hardlo and Faten Ashour	6265
<i>Norm It! Lexical Normalization for Italian and Its Downstream Effects for Dependency Parsing</i> Rob van der Goot, Alan Ramponi, Tommaso Caselli, Michele Cafagna and Lorenzo De Mattei	6272
<i>TARc: Incrementally and Semi-Automatically Collecting a Tunisian Arabish Corpus</i> Elisa Gugliotta and Marco Dinarelli	6279
<i>Small Town or Metropolis? Analyzing the Relationship between Population Size and Language</i> Amy Rechkemmer, Steven Wilson and Rada Mihalcea	6287

<i>Inferring Social Media Users' Mental Health Status from Multimodal Information</i> Zhentao Xu, Verónica Pérez-Rosas and Rada Mihalcea	6292
<i>Synthetic Data for English Lexical Normalization: How Close Can We Get to Manually Annotated Data?</i> Kelly Dekker and Rob van der Goot	6300
<i>A Corpus of German Reddit Exchanges (GeRedE)</i> Andreas Blombach, Natalie Dykes, Philipp Heinrich, Besim Kabashi and Thomas Proisl	6310
<i>French Tweet Corpus for Automatic Stance Detection</i> Marc Evrard, Rémi Uro, Nicolas Hervé and Béatrice Mazoyer	6317
<i>LSCP: Enhanced Large Scale Colloquial Persian Language Understanding</i> Hadi Abdi Khojasteh, Ebrahim Ansari and Mahdi Bohlouli	6323
Speech Recognition and Synthesis	
<i>Burmese Speech Corpus, Finite-State Text Normalization and Pronunciation Grammars with an Application to Text-to-Speech</i> Yin May Oo, Theeraphol Wattanavekin, Chenfang Li, Pasindu De Silva, Supheakmungkol Sarin, Knot Pipatsrisawat, Martin Jansche, Oddur Kjartansson and Alexander Gutkin	6328
<i>Evaluating and Improving Child-Directed Automatic Speech Recognition</i> Eric Booth, Jake Carns, Casey Kennington and Nader Rafla	6340
<i>Parallel Corpus for Japanese Spoken-to-Written Style Conversion</i> Mana Ihori, Akihiko Takashima and Ryo Masumura	6346
<i>Multi-Stage Cross-Lingual Acoustic Model Adaption for Robust Speech Recognition in Real-World Applications - A Case Study on German Oral History Interviews</i> Michael Gref, Oliver Walter, Christoph Schmidt, Sven Behnke and Joachim Köhler	6354
<i>Large Corpus of Czech Parliament Plenary Hearings</i> Jonas Kratochvil, Peter Polak and Ondrej Bojar	6363
<i>Augmented Prompt Selection for Evaluation of Spontaneous Speech Synthesis</i> Eva Szekely, Jens Edlund and joakim gustafson	6368
<i>ATC-ANNO: Semantic Annotation for Air Traffic Control with Assistive Auto-Annotation</i> Marc Schulder, Johannah O'Mahony, Yury Bakanouski and Dietrich Klakow	6375
<i>MASRI-HEADSET: A Maltese Corpus for Speech Recognition</i> Carlos Daniel Hernandez Mena, Albert Gatt, Andrea DeMarco, Claudia Borg, Lonneke van der Plas, Amanda Muscat and Ian Padovani	6381
<i>Automatic Period Segmentation of Oral French</i> Natalia Kalashnikova, Loïc Grobol, Iris Eshkol-Taravella and François Delafontaine	6389
<i>Corpus Generation for Voice Command in Smart Home and the Effect of Speech Synthesis on End-to-End SLU</i> Thierry Desot, François Portet and Michel Vacher	6395
<i>Text and Speech-based Tunisian Arabic Sub-Dialects Identification</i> Najla Ben Abdallah, Saméh Kchaou and Fethi Bougares	6405

<i>Urdu Pitch Accents and Intonation Patterns in Spontaneous Conversational Speech</i> Luca Rognoni, Judith Bishop, Miriam Corris, Jessica Fernando and Rosanna Smith	6412
<i>IndicSpeech: Text-to-Speech Corpus for Indian Languages</i> Nimisha Srivastava, Rudrabha Mukhopadhyay, Prajwal K R and C V Jawahar	6417
<i>Using Automatic Speech Recognition in Spoken Corpus Curation</i> Jan Gorisch, Michael Gref and Thomas Schmidt	6423
<i>Integrating Disfluency-based and Prosodic Features with Acoustics in Automatic Fluency Evaluation of Spontaneous Speech</i> Huaijin Deng, Youchao Lin, Takehito Utsuro, Akio Kobayashi, Hiromitsu Nishizaki and Junichi Hoshino	6429
<i>DNN-based Speech Synthesis Using Abundant Tags of Spontaneous Speech Corpus</i> Yuki Yamashita, Tomoki Koriyama, Yuki Saito, Shinnosuke Takamichi, Yusuke Ijima, Ryo Masumura and Hiroshi Saruwatari	6438
<i>Automatic Speech Recognition for Uyghur through Multilingual Acoustic Modeling</i> Ayimunishagu Abulimiti and Tanja Schultz	6444
<i>The SAFE-T Corpus: A New Resource for Simulated Public Safety Communications</i> Dana Delgado, Kevin Walker, Stephanie Strassel, Karen Jones, Christopher Caruso and David Graff 6450	
<i>Lexical Tone Recognition in Mizo using Acoustic-Prosodic Features</i> Parismita Gogoi, Abhishek Dey, Wendy Lalhminghlui, Priyankoo Sarmah and S R Mahadeva Prasanna	6458
<i>Artie Bias Corpus: An Open Dataset for Detecting Demographic Bias in Speech Applications</i> Josh Meyer, Lindy Rauchenstein, Joshua D. Eisenberg and Nicholas Howell	6462
<i>Evaluation of Off-the-shelf Speech Recognizers Across Diverse Dialogue Domains</i> Kallirroi Georgila, Anton Leuski, Volodymyr Yanov and David Traum	6469
<i>CEASR: A Corpus for Evaluating Automatic Speech Recognition</i> Malgorzata Anna Ulasik, Manuela Hürlimann, Fabian Germann, Esin Gedik, Fernando Benites and Mark Cieliebak	6477

Speech Resource/Database

<i>MaSS: A Large and Clean Multilingual Corpus of Sentence-aligned Spoken Utterances Extracted from the Bible</i> Marcely Zanon Boito, William Havard, Mahault Garnerin, Éric Le Ferrand and Laurent Besacier	6486
<i>Open-source Multi-speaker Speech Corpora for Building Gujarati, Kannada, Malayalam, Marathi, Tamil and Telugu Speech Synthesis Systems</i> Fei He, Shan-Hui Cathy Chu, Oddur Kjartansson, Clara Rivera, Anna Katanova, Alexander Gutkin, Isin Demirsahin, Cibu Johny, Martin Jansche, Supheakmungkol Sarin and Knot Pipatsrisawat	6494

<i>Crowdsourcing Latin American Spanish for Low-Resource Text-to-Speech</i> Adriana Guevara-Rukoz, Isin Demirsahin, Fei He, Shan-Hui Cathy Chu, Supheakmungkol Sarin, Knot Pipatsrisawat, Alexander Gutkin, Alena Butryna and Oddur Kjartansson	6504
<i>A Manually Annotated Resource for the Investigation of Nasal Grunts</i> Aurélie Chlébowski and Nicolas Ballier	6514
<i>The Objective and Subjective Sleepiness Voice Corpora</i> Vincent P. Martin, Jean-Luc Rouas, Jean-Arthur Micoulaud Franchi and Pierre Philip	6523
<i>Open-source Multi-speaker Corpora of the English Accents in the British Isles</i> Isin Demirsahin, Oddur Kjartansson, Alexander Gutkin and Clara Rivera	6532
<i>TV-AfD: An Imperative-Annotated Corpus from The Big Bang Theory and Wikipedia’s Articles for Deletion Discussions</i> Yimin Xiao, Zong-Ying Slaton and Lu Xiao	6542
<i>A Large Scale Speech Sentiment Corpus</i> Eric Chen, Zhiyun Lu, Hao Xu, Liangliang Cao, Yu Zhang and James Fan	6549
<i>SibLing Corpus of Russian Dialogue Speech Designed for Research on Speech Entrainment</i> Tatiana Kachkovskaia, Tatiana Chukaeva, Vera Evdokimova, Pavel Kholiavin, Natalia Kriakina, Daniil Kocharov, Anna Mamushina, Alla Menshikova and Svetlana Zimina	6556
<i>PhonBank and Data Sharing: Recent Developments in European Portuguese</i> Ana Margarida Ramalho, Maria João Freitas and Yvan Rose	6562
<i>SMASH Corpus: A Spontaneous Speech Corpus Recording Third-person Audio Commentaries on Gameplay</i> Yuki Saito, Shinnosuke Takamichi and Hiroshi Saruwatari	6571
<i>Improving Speech Recognition for the Elderly: A New Corpus of Elderly Japanese Speech and Investigation of Acoustic Modeling for Speech Recognition</i> Meiko Fukuda, Hiromitsu Nishizaki, Yurie Iribe, Ryota Nishimura and Norihide Kitaoka	6578
<i>Preparation of Bangla Speech Corpus from Publicly Available Audio & Text</i> Shafayat Ahmed, Nafis Sadeq, Sudipta Saha Shubha, Md. Nahidul Islam, Muhammad Abdullah Adnan and Mohammad Zuberul Islam	6586
<i>On Construction of the ASR-oriented Indian English Pronunciation Dictionary</i> Xian Huang, Xin Jin, Qike Li and Keliang Zhang	6593
<i>Gender Representation in Open Source Speech Resources</i> Mahault Garnerin, Solange Rossato and Laurent Besacier	6599
<i>RSC: A Romanian Read Speech Corpus for Automatic Speech Recognition</i> Alexandru-Lucian Georgescu, Horia Cucu, Andi Buzo and Corneliu Burileanu	6606
<i>FAB: The French Absolute Beginner Corpus for Pronunciation Training</i> Sean Robertson, Cosmin Munteanu and Gerald Penn	6613
<i>Call My Net 2: A New Resource for Speaker Recognition</i> Karen Jones, Stephanie Strassel, Kevin Walker and Jonathan Wright	6621
<i>DaCToR: A Data Collection Tool for the RELATER Project</i> Juan Hussain, Oussama Zenkri, Sebastian Stüker and Alex Waibel	6627

<i>Development and Evaluation of Speech Synthesis Corpora for Latvian</i>	
Roberts Dargis, Peteris Paikens, Normunds Gruzitis, Ilze Auzina and Agate Akmane	6633

Summarisation

<i>Abstractive Document Summarization without Parallel Data</i>	
Nikola I. Nikolov and Richard Hahnloser	6638
<i>GameWikiSum: a Novel Large Multi-Document Summarization Dataset</i>	
Diego Antognini and Boi Faltings	6645
<i>Summarization Corpora of Wikipedia Articles</i>	
Dominik Frefel	6651
<i>Language Agnostic Automatic Summarization Evaluation</i>	
Christopher Tauchmann and Margot Mieskes	6656
<i>Two Huge Title and Keyword Generation Corpora of Research Articles</i>	
Erion Çano and Ondřej Bojar	6663
<i>A Multi-level Annotated Corpus of Scientific Papers for Scientific Document Summarization and Cross-document Relation Discovery</i>	
Ahmed AbuRa'ed, Horacio Saggion and Luis Chiruzzo	6672
<i>Abstractive Text Summarization based on Language Model Conditioning and Locality Modeling</i>	
Dmitrii Aksenov, Julian Moreno-Schneider, Peter Bourgonje, Robert Schwarzenberg, Leonhard Hennig and Georg Rehm	6680
<i>A Data Set for the Analysis of Text Quality Dimensions in Summarization Evaluation</i>	
Margot Mieskes, Eneldo Loza Mencía and Tim Kronsbein	6690
<i>Summarization Beyond News: The Automatically Acquired Fandom Corpora</i>	
Benjamin Hättasch, Nadja Geisler, Christian M. Meyer and Carsten Binnig	6700
<i>Invisible to People but not to Machines: Evaluation of Style-aware Headline Generation in Absence of Reliable Human Judgment</i>	
Lorenzo De Mattei, Michele Cafagna, Felice Dell'Orletta and Malvina Nissim	6709
<i>Align then Summarize: Automatic Alignment Methods for Summarization Corpus Creation</i>	
Paul Tardy, David Janiszek, Yannick Estève and Vincent Nguyen	6718
<i>A Summarization Dataset of Slovak News Articles</i>	
Marek Suppa and Jergus Adamec	6725
<i>DaNewsroom: A Large-scale Danish Summarisation Dataset</i>	
Daniel Varab and Natalie Schluter	6731

Text Mining

<i>Diverging Divergences: Examining Variants of Jensen Shannon Divergence for Corpus Comparison Tasks</i>	
Jinghui Lu, Maeve Henchion and Brian Mac Namee	6740

<i>TopicNet: Making Additive Regularisation for Topic Modelling Accessible</i>	
Victor Bulatov, Vasiliy Alekseev, Konstantin Vorontsov, Darya Polyudova, Eugenia Veselova, Alexey Goncharov and Evgeny Egorov	6745
<i>SC-CoMics: A Superconductivity Corpus for Materials Informatics</i>	
Kyosuke Yamaguchi, Ryoji Asahi and Yutaka Sasaki	6753
<i>GitHub Typo Corpus: A Large-Scale Multilingual Dataset of Misspellings and Grammatical Errors</i>	
Masato Hagiwara and Masato Mita	6761
<i>Annotation of Adverse Drug Reactions in Patients' Weblogs</i>	
Yuki Arase, Tomoyuki Kajiwarara and Chenhui Chu	6769
<i>Beyond Citations: Corpus-based Methods for Detecting the Impact of Research Outcomes on Society</i>	
Rezvaneh Rezapour, Jutta Bopp, Norman Fiedler, Diana Steffen, Andreas Witt and Jana Diesner	6777
<i>Toxic, Hateful, Offensive or Abusive? What Are We Really Classifying? An Empirical Analysis of Hate Speech Datasets</i>	
Paula Fortuna, Juan Soler and Leo Wanner	6786
<i>Unsupervised Argumentation Mining in Student Essays</i>	
Isaac Persing and Vincent Ng	6795
<i>Aspect-Based Sentiment Analysis as Fine-Grained Opinion Mining</i>	
Gerardo Ocampo Diaz, Xuanming Zhang and Vincent Ng	6804
<i>Predicting Item Survival for Multiple Choice Questions in a High-Stakes Medical Exam</i>	
Victoria Yaneva, Le An Ha, Peter Baldwin and Janet Mee	6812
Textual Entailment and Paraphrasing	
<i>Discourse Component to Sentence (DC2S): An Efficient Human-Aided Construction of Paraphrase and Sentence Similarity Dataset</i>	
Won Ik Cho, Jong In Kim, Young Ki Moon and Nam Soo Kim	6819
<i>Japanese Realistic Textual Entailment Corpus</i>	
Yuta Hayashibe	6827
<i>Improving the Precision of Natural Textual Entailment Problem Datasets</i>	
Jean-Philippe Bernardy and Stergios Chatzikyriakidis	6835
<i>Comparative Study of Sentence Embeddings for Contextual Paraphrasing</i>	
Louisa Pragst, Wolfgang Minker and Stefan Ultes	6841
<i>HypoNLI: Exploring the Artificial Patterns of Hypothesis-only Bias in Natural Language Inference</i>	
Tianyu Liu, Zheng Xin, Baobao Chang and Zhifang Sui	6852
<i>SAPPHERE: Simple Aligner for Phrasal Paraphrase with Hierarchical Representation</i>	
Masato Yoshinaka, Tomoyuki Kajiwarara and Yuki Arase	6861
<i>TaPaCo: A Corpus of Sentential Paraphrases for 73 Languages</i>	
Yves Scherrer	6868

<i>Automated Fact-Checking of Claims from Wikipedia</i>	
Aalok Sathe, Salar Ather, Tuan Manh Le, Nathan Perry and Joonsuk Park	6874
<i>Towards the Necessity for Debiasing Natural Language Inference Datasets</i>	
Mithun Paul Panenghat, Sandeep Suntwal, Faiz Rafique, Rebecca Sharp and Mihai Surdeanu	6883
<i>A French Corpus for Semantic Similarity</i>	
Rémi Cardon and Natalia Grabar	6889
<i>Developing Dataset of Japanese Slot Filling Quizzes Designed for Evaluation of Machine Reading Comprehension</i>	
Takuto Watarai and Masatoshi Tsuchiya	6895
Tools, Systems, Applications	
<i>Detecting Negation Cues and Scopes in Spanish</i>	
Salud María Jiménez-Zafra, Roser Morante, Eduardo Blanco, María Teresa Martín Valdivia and L. Alfonso Ureña López	6902
<i>TIARA: A Tool for Annotating Discourse Relations and Sentence Reordering</i>	
Jan Wira Gotama Putra, Simone Teufel, Kana Matsumura and Takenobu Tokunaga	6912
<i>Infrastructure for Semantic Annotation in the Genomics Domain</i>	
Mahmoud El-Haj, Nathan Rutherford, Matthew Coole, Ignatius Ezeani, Sheryl Prentice, Nancy Ide, Jo Knight, Scott Piao, John Mariani, Paul Rayson and Keith Suderman	6921
<i>Correcting the Autocorrect: Context-Aware Typographical Error Correction via Training Data Augmentation</i>	
Kshitij Shah and Gerard de Melo	6930
<i>KidSpell: A Child-Oriented, Rule-Based, Phonetic Spellchecker</i>	
Brody Downs, Oghenemaro Anuyah, Aprajita Shukla, Jerry Alan Fails, Sole Pera, Katherine Wright and Casey Kennington	6937
<i>ThaiLMCut: Unsupervised Pretraining for Thai Word Segmentation</i>	
Suteera Seeha, Ivan Bilan, Liliana Mamani Sanchez, Johannes Huber, Michael Matuschek and Hinrich Schütze	6947
<i>CCOHA: Clean Corpus of Historical American English</i>	
Reem Alatrash, Dominik Schlechtweg, Jonas Kuhn and Sabine Schulte im Walde	6958
<i>Outbound Translation User Interface Ptakopět: A Pilot Study</i>	
Vilém Zouhar and Ondřej Bojar	6967
<i>Seshat: a Tool for Managing and Verifying Annotation Campaigns of Audio Data</i>	
Hadrien Titeux, Rachid Riad, Xuan-Nga Cao, Nicolas Hamilakis, Kris Madden, Alejandrina Cristia, Anne-Catherine Bachoud-Lévi and Emmanuel Dupoux	6976
<i>Dragonfly: Advances in Non-Speaker Annotation for Low Resource Languages</i>	
Cash Costello, Shelby Anderson, Caitlyn Bishop, James Mayfield and Paul McNamee	6983
<i>Natural Language Processing Pipeline to Annotate Bulgarian Legislative Documents</i>	
Svetla Koeva, Nikola Obreshkov and Martin Yalamov	6988

<i>CLDFBench: Give Your Cross-Linguistic Data a Lift</i> Robert Forkel and Johann-Mattis List	6995
<i>KonText: Advanced and Flexible Corpus Query Interface</i> Tomáš Machálek	7003
<i>Word at a Glance: Modular Word Profile Aggregator</i> Tomáš Machálek	7009
<i>RKorAPClient: An R Package for Accessing the German Reference Corpus DeReKo via KorAP</i> Marc Kupietz, Nils Diewald and Eliza Margaretha	7015
<i>CAMEL Tools: An Open Source Python Toolkit for Arabic Natural Language Processing</i> Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhil Eryani, Alexander Erdmann and Nizar Habash	7022
<i>ReSiPC: a Tool for Complex Searches in Parallel Corpora</i> Antoni Oliver and Bojana Mikelenić	7033
<i>HitzalMed: Anonymisation of Clinical Text in Spanish</i> Salvador Lima Lopez, Naiara Perez, Laura García-Sardiña and Montse Cuadros	7038
<i>The xtsv Framework and the Twelve Virtues of Pipelines</i> Balázs Indig, Bálint Sass and Iván Mittelholz	7044
<i>A Web-based Collaborative Annotation and Consolidation Tool</i> Tobias Daudert	7053
<i>Data Query Language and Corpus Tools for Slot-Filling and Intent Classification Data</i> Stefan Larson, Eric Guldán and Kevin Leach	7060
<i>SHR++: An Interface for Morpho-syntactic Annotation of Sanskrit Corpora</i> Amrith Krishna, Shiv Vidhyut, Dilpreet Chawla, Sruti Sambhavi and Pawan Goyal	7069
<i>KOTONOHA: A Corpus Concordance System for Skewer-Searching NINJAL Corpora</i> Teruaki Oka, Yuichi Ishimoto, Yutaka Yagi, Takenori Nakamura, Masayuki Asahara, Kikuo Maekawa, Toshinobu Ogiso, Hanae Koiso, Kumiko Sakoda and Nobuko Kibe	7077
<i>Gamification Platform for Collecting Task-oriented Dialogue Data</i> Haruna Ogawa, Hitoshi Nishikawa, Takenobu Tokunaga and Hikaru Yokono	7084
<i>Improving the Production Efficiency and Well-formedness of Automatically-Generated Multiple-Choice Cloze Vocabulary Questions</i> Ralph Rose	7094
<i>Improving Sentence Boundary Detection for Spoken Language Transcripts</i> Ines Rehbein, Josef Ruppenhofer and Thomas Schmidt	7102
<i>MorphAGram, Evaluation and Framework for Unsupervised Morphological Segmentation</i> Ramy Eskander, Francesca Callejas, Elizabeth Nichols, Judith Klavans and Smaranda Muresan	7112
<i>CTAP for Italian: Integrating Components for the Analysis of Italian into a Multilingual Linguistic Complexity Analysis Tool</i> Nadezda Okinina, Jennifer-Carmen Frey and Zarah Weiss	7123

<i>Do you Feel Certain about your Annotation? A Web-based Semantic Frame Annotation Tool Considering Annotators' Concerns and Behaviors</i>	
Regina Stodden, Behrang QasemiZadeh and Laura Kallmeyer	7132
<i>Seq2SeqPy: A Lightweight and Customizable Toolkit for Neural Sequence-to-Sequence Modeling</i>	
Raheel Qader, François Portet and Cyril Labbe	7140
<i>Profiling-UD: a Tool for Linguistic Profiling of Texts</i>	
Dominique Brunato, Andrea Cimino, Felice Dell'Orletta, Giulia Venturi and Simonetta Montemagni	7145
<i>EstNLTK 1.6: Remastered Estonian NLP Pipeline</i>	
Sven Laur, Siim Orasmaa, Dage Särg and Paul Tammo	7152
<i>A Tree Extension for CoNLL-RDF</i>	
Christian Chiarcos and Luis Glaser	7161
<i>Lemmatising Verbs in Middle English Corpora: The Benefit of Enriching the Penn-Helsinki Parsed Corpus of Middle English 2 (PPCME2), the Parsed Corpus of Middle English Poetry (PCMEP), and A Parsed Linguistic Atlas of Early Middle English (PLAEME)</i>	
Carola Trips and Michael Percillier	7170
<i>CoCo: A Tool for Automatically Assessing Conceptual Complexity of Texts</i>	
Sanja Stajner, Sergiu Nisioi and Ioana Hulpuş	7179
<i>PyVallex: A Processing System for Valency Lexicon Data</i>	
Jonathan Verner and Anna Vernerová	7187
<i>Editing OntoLex-Lemon in VocBench 3</i>	
Manuel Fiorelli, Armando Stellato, Tiziano Lorenzetti, Andrea Turbati, Peter Schmitz, Enrico Francesconi, Najeh Hajlaoui and Brahim Batouche	7194
<i>MALT-IT2: A New Resource to Measure Text Difficulty in Light of CEFR Levels for Italian L2 Learning</i>	
Luciana Forti, Giuliana Grego Bolli, Filippo Santarelli, Valentino Santucci and Stefania Spina	7204
<i>Fintan - Flexible, Integrated Transformation and Annotation eEngineering</i>	
Christian Fäth, Christian Chiarcos, Björn Ebbrecht and Maxim Ionov	7212
<i>Contemplata, a Free Platform for Constituency Treebank Annotation</i>	
Jakub Waszczuk, Ilaine Wang, Jean-Yves Antoine and Anaïs Halftermeyer	7222
<i>Interchange Formats for Visualization: LIF and MMIF</i>	
Kyeongmin Rim, Kelley Lynch, Marc Verhagen, Nancy Ide and James Pustejovsky	7230
<i>Developing NLP Tools with a New Corpus of Learner Spanish</i>	
Sam Davidson, Aaron Yamada, Paloma Fernandez Mira, Agustina Carando, Claudia H. Sanchez Gutierrez and Kenji Sagae	7238
<i>DeepNLPF: A Framework for Integrating Third Party NLP Tools</i>	
Francisco Rodrigues, Rinaldo Lima, William Domingues, Robson Fidalgo, Adrian Chifu, Bernard Espinasse and Sébastien Fournier	7244