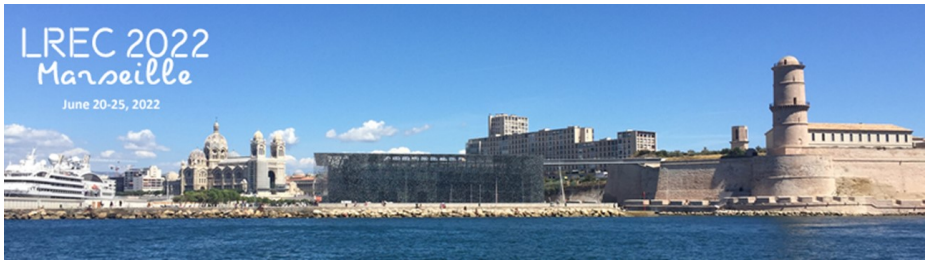


Language Resources and Evaluation Conference
LREC 2022
20-25 June 2022

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, Jan Odijk, Stelios Piperidis

# LREC 2022, THIRTEEN INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION

## LREC 2022 Conference Proceedings

**ISBN: 979-10-95546-72-6**
**EAN: 9791095546726**

**For more information:**
European Language Resources Association (ELRA)
9 rue des Cordelières
75013, Paris
France
http://www.elra.info
Email: lrec@elda.org

© European Language Resources Association (ELRA)

These LREC2022 proceedings are licensed under a Creative Commons
Attribution-NonCommercial 4.0 International License

# LREC 2022

# 13th International Conference on Language Resources and Evaluation

## Introduction by Nicoletta Calzolari

## Chair of the 13th edition of LREC
## ELRA Honorary President

**Introduction to LREC2022**
Message from the LREC 2022 Conference Chair

Welcome to the $13^{th}$ edition of LREC in Marseille, first time an LREC in France! It will be a big emotion to welcome so many of you at the Opening session, finally meeting again so many colleagues and friends!

We will enjoy together not only the conference but also the special atmosphere of Marseille and the wonderful view of the Mediterranean and the city from the Palais du Pharo.

The previous LREC, that didn't happen, would have been an occasion to celebrate the $25^{th}$ anniversary of ELRA. We will celebrate this time the next 25 years. We are proud that ELRA is a mature association. And LREC too: LREC started in 1998, 24 years ago, and now it is one of the most successful conferences of the field.

***A major novelty in the organisation of LREC: the involvement of many colleagues.***

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

This time we decided, in the ELRA Board and the LREC Program Committee, it was time to start changing the way in which LREC is organised and managed. In addition to the usual Program Committee we decided to involve in the management and decision process many colleagues, as is done in all major conferences.

We thus recruited 3 Workshop chairs, 2 Tutorial chairs and for the main conference 69 Area chairs. It was important for me to try to convey to them the "spirit" of LREC. LREC seeks to: increase representation among under-resourced languages, value all types of contributors to the field (corpus creators, algorithm developers, technology evaluators, program managers), reduce barriers to new participants, encourage meaningful scientific participation among industrials, and otherwise provide opportunities to those who might otherwise be silenced. To pay proper attention to these issues is a distinguishing feature of LREC.

Area chairs' suggestions about acceptance were then analysed by the Program Committee. I was acting in a double role, as General chair and Program chair, a big task (in about 3 months I exchanged almost

2700 mails mainly with Area chairs) and a big responsibility.

Continuing in the change, I think that next time we must also have 3 Program chairs.

***Some LREC2022 figures***

2022 figures are similar to those of 2020.

We received 50 Workshop proposals, and 12 Tutorials. Out of these 38 Workshops and 9 Tutorials were accepted.

We received 1302 submissions, of which 45 were Desk rejected or Withdrawn before reviewing. We thus had 1257 submissions to review.

A very large part of our community was involved in the reviewing effort, to be able to assign few papers per reviewer. We recruited 1628 colleagues to be part of the Scientific Committee, out of which 1401 were assigned submissions to review. Many more were invited but many refused or didn't answer. Finding reviewers is becoming a quite difficult task, good reviewers even more.

Few reviewers did not complete the task. For this I asked the Area chairs to have some emergency reviewers.

Area chairs analysed more than 3845 reviews (some tracks had 4 reviewers per paper) and made suggestions of acceptance/rejection. After that the LREC Program Committee arrived at the final decisions in 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 now in the program 804 papers, of which 156 onsite Orals, 416 onsite Posters, and 232 Remote.

We have broken the record of registrations, almost 1600 in this moment (with 942 attending in person, quite many), probably more at the end.

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

***LREC2022 as a hybrid event***

Another major change is that we decided to organise LREC2022 as a hybrid event.

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. For all these reasons we, as Program Committee, decided to try to privilege onsite participation. But obviously because of Covid we had to allow also for remote participation.

Most probably this is a change that will remain in the future, also after Covid. I think we must reassess how to organise big conferences. Allowing for remote participation will make our conferences more inclusive, allowing participation also to those who for any reason are not able to travel.

Because of this this year in addition to the usual Oral and Poster sessions the Program is composed also or the Remote only sessions.

I must say that the organisation of a hybrid conference is much more demanding, not only for the local organisation but also for the scientific aspects. Until the last moment in fact there have been, as expected, changes in attendance (in person or remote, in both directions) which caused quite a number of changes in the organisation of the programs.

***LREC2022 Proceedings***

Because of the remote participation, also the Proceedings are organised slightly differently. The Proceedings in fact reflect the program, with the usual in person Oral and Poster sessions, following the timetable of the Program.

This time they are followed by the Remote only sessions, organised on the basis of the 22 Tracks. This Remote sessions contain remote Orals and Posters together: this is in line with what we always say, that the distinction between oral and poster 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.

As we did last time, we decided to print 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.

***LREC2022 acceptance rate: a motivated choice for an inclusive conference***

The LREC2022 acceptance rate is 61.75%, 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 values topics – such as Less-resourced languages or Infrastructural issues, strategies and policies – that may not easily find proper venues in other big conferences. Research is strongly affected also by infrastructural

(meta-research) activities, really needed for our field to progress.

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 2021 Google Scholar Metrics h5-index, LREC ranks $6^{th}$ of the big conferences/journals in Computational Linguistics.

***More popular topics in LREC2022***

This time we had 22 Areas and authors were asked to select a main and a secondary. Submissions to the areas gives us an indication of the current interests of our participants.

By large the most popular areas (more than 100 submissions) are:

- Corpora and Annotation (including Tools, Systems, Treebanks)
- Information Extraction and Information Retrieval (including NER, QA, Text Mining, Document Classification, Text Categorisation)
- Areas with 50 or more submissions are:
- Applications involving LRs and Evaluation (including applications in specific domains)
- Less-Resourced/Endangered Languages
- Dialogue, Conversational Systems, Chatbots, Human-Robot Interaction
- Speech Resources and Processing (including Phonetic Databases, Phonology, Prosody)
- Statistical Methods and Machine Learning for Language Technologies (including Language Models)
- Multilinguality and Machine Translation (including Speech-to-Speech translation)
- Evaluation and Validation Methodologies
- Opinion Mining, Sentiment Analysis, Emotion Recognition/Generation
- Multimodality and Cross-modality (including Sign Languages, Vision and other modalities) and Multimedia
- Social Media Processing
- Statistical Methods and Machine Learning for Language Technologies (including Language Models)

While the areas that this time attracted less attention are:

- Natural Language Generation (including Summarization)

- Lexicons (also WordNet, FrameNet, Multimodal and Sign Language lexicons, etc.)
- Semantics (including Distributional Semantics, Word Sense Disambiguation, Coreference, etc.)
- Language Resources and Evaluation for Psycho-linguistics, Cognitive Linguistics and Linguistic Theories
- Digital Humanities and Cultural Heritage
- Discourse and Pragmatics
- Parsing, Tagging, Grammar, Syntax, Morphology
- Language Resource Infrastructures, Standards for LRs, Metadata, Policy issues, Ethics, Legal Issues
- Reproduction of Research Results in Science and Technology of Language

***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 made this LREC2022 possible.

I express my deepest gratitude to the new Chairs that we introduced this time: to the Workshop and Tutorial chairs and to the Area chairs. They all did a tremendous job.

As usual, I thank the old time friends of the Programme Committee: with them we went through the suggestions of the other chairs to make the final decisions. 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, Jan Odijk, Stelios Piperidis. A particular thanks goes to Thierry Declerk, who helped me in the preparation of the program, and to Khalid Choukri, who has worked for the preparation of the Proceedings in START. I am very grateful to Frédéric Béchet and Philippe Blache, chairs of the Local Committee, who dealt not only with local matters but also with matters of the hybrid organisation.

I thank the ELRA Association: LREC is a major service from ELRA to all the community. The ELRA Board members contributed to the change in the organisation and helped also in suggesting names of chairs for the various Areas.

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 Hélène Mazo, and I thank them for their dedication and competence in managing the many tasks they have in their hands. And I add my thanks to many other colleagues of the two institutions who contributed in different ways, and in particular to Victoria Arranz, Roberto Bartolini, Lucille Blanchard, Francesca Frontini, Valérie Mapelli, Monica Monachini, Vincenzo Parrinelli, Valeria Quochi, Caroline Rannaud, Mickaël Rigaud, Alexandre Sicard, Kamelia Yahiaoui.

I express my deep gratitude to the local authorities and to the sponsors that have generously 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, who enrich LREC with so many interesting events. I hope the hybrid organisation will work with them too.

I thank the Catalyst company that helped us in the organisation of the hybrid event.

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 am looking forward meeting many of you in Marseille.

And this time again I can thank all the LREC participants, both the remote and the onsite ones. I am so glad to meet so many of you soon. I really hope this LREC will open new research paths and new exciting work for many of you, also profiting of the many contacts you will have.

Marseille, with its special light, will be a fantastic Mediterranean location for LREC2022.

I welcome you at LREC2022, and with all the Programme Committee I hope you have fruitful contacts and find interesting and valuable presentations at the conference and readings in the LREC2022 Proceedings.

Enjoy LREC2022 in Marseille!

Nicoletta Calzolari

Chair of the $13^{th}$ International Conference on
Language Resources and Evaluation

and ELRA Honorary President

# LREC 2022

# 13th International Conference on Language Resources and Evaluation

## Opening Address

António Branco, President of ELRA

# LREC 2022

# 13th International Conference on Language Resources and Evaluation

# Opening Address

## ELRA Secretary General and ELDA Chief Executive Officer
## Khalid Choukri

Dear guests, dear friends, Dear ELRA members, Dear LREC Participants,
Je suis très heureux que nous ayons repris la vie que nous chérissons tous en nous retrouvons ici à Marseille tous ensemble pour LREC 2022, avec l'espoir que la pandémie ne vous a pas trop impacté.

It is my great pleasure to welcome you all, four years after our last face to face LREC conference in Miyazaki, Japan, in 2018.

Welcome to all those who joined us today here in Marseille but also warm greetings to those who could not join us in person but are participating remotely.

It has been a challenge to organize an LREC as a hybrid conference, LREC which spirit is to meet friends and colleagues and revitalize our networks. This new framework (in person and remote participation) requires strict organizational processes and logistics and we hope that we will succeed to make the best out of it.

Welcome to this 13th edition of LREC.

## ELRA & ELDA missions

ELRA has gone trough an important restructuring to better account for the community expectations in terms of Language Resources and Evaluation of technologies. Some of this is elaborated upon by our president in his message. The major changes are related to the association membership and its governance. The technical, legal, logistics and other practical tasks will continue to be our core business while improving our connections to our members.

As most of you know, ELRA missioned ELDA to carry out its operational tasks and to put in operation its strategy. The General Assembly decides on the annual actions to conduct, and regular strategic meetings design the long-term perspectives and the associated roadmaps.

I have been proud to coordinate such activities with an excellent multidisciplinary team. Let me

highlight some of the practical tasks that we managed sustainably for over a quarter of a century.

## LREC to promote recognition of LRs & Eval research

I am sure that many of you still remember the first time, 1998, in Granada after that ELRA had decided to establish a conference, fully dedicated to Language Resources and Evaluation of Human Language Technologies. ELRA focused on topics the were and still are fundamentally relevant to ensure that researchers can devote the necessary efforts to produce and package or re-purpose Language Resources and get the deserved peers recognition. We all understand that such recognition is essential for the curricula, through the citations and conference indexing. From the start, the idea was to oversee these topics in the largest inclusive context possible. It was of paramount importance to promote the field beyond the main streams of the moment. The main streams were mostly a strong focus on languages such as English, Mandarin, Arabic and very few European languages and with very few technical approaches.

LREC genesis came with all languages and modalities in its DNA and hence promoted works on less-resourced languages and in particular the indigenous ones spoken by small communities. They also need to benefit from our achievements. Since that first LREC and the promotion of language resources for all languages, we can see the smart increase in the number of languages that are tackled by the research and industry communities. LREC shows this through the number of languages cited in the LREMap (`https://lremap.elra.info/`) . LREMap, as initiated by ELRA at LREC2010, help report on the language resources used/produced/used by the LREC authors. The initiative has been taken by other major conferences today. With about 8000 items and more than 100 different languages today, LREMap emphasizes the progress made so far. Nevertheless, we are still far away from the 7000 human languages out of which at least 30% would positively benefit from our field work.

The other important dimension of LREC (and LRE, Language Resources and Evaluation Journal, the associated ELRA Journal by Springer) is the promotion of evaluation of technologies and reproducibility of experiments. The number of evaluation campaigns has grown exponentially and, while this is necessary for the assessment of existing approaches, it became impossible to keep track and monitor this through an accessible inventory. It became so easy to initiate a challenge but very difficult to design this in a long-term roadmap with strong incentives that federate a large community able to conduct reproducible research. IT is very common to see a challenge established for our round without too much scientific conclusions for the future. Some initiatives started focusing on less resourced languages. This should be highly encouraged and may be made mandatory so each time a benchmarking is initiated, at least one language that is not part of the few large ones is considered if it makes sense. Of course, the budget may not always be sufficient but many funding agencies would be ready to support such action. ELRA joined forces with other partners and is willing to continue such support if necessary.

It is crucial that the community decides to coordinate such challenges and set-up a repository for the campaigns but also ensure that developed resources, metrics, know-how is documented and made available. ELRA is prepared to play a role on this.

## ELRA Catalogues & ISLRN

Our primary mission was to act as a data center for Language Resources sharing. To this end, we devoted our efforts to identifying useful data sets, reviewing the community publications, the outcome of the funded projects and the disseminated information. We worked on clearing all Intellectual Property Rights (IPR) and other legal issues with the right holders (not necessarily the data owners). Additional efforts were needed to document them using very detailed meta-data descriptors and we are happy to see these meta-data schemas re-used by others. We also designed the appropriate licenses to make the LRs easily licensable and usable for various purposes. Licensing remains a critical task as we need to balance the expectations of all parties (academic users, commercial partners, funding agencies, right holders, etc.). These different tasks, important to ensure the possibility to share LRs, are documented in our Data Management Plan (DMP, `http://www.elra.info/en/services-around-lrs/dmp/`) that we make available and customizable by third parties for their own resources with our free support when needed.

We also included, in our identification and cataloguing work of LRs, the packages produced within the evaluation campaigns I mentioned above. The package often comprises the test data, the metrics, as well as the benchmarking results and publications. These packages allow to reproduce the work done with an evaluation campaign by newcomers and consist of a useful "Exist strategy" that capitalizes on the work carried out during the campaign.

As of today, over 1500 Language resources are catalogued by ELRA and made available, covering more than 80 languages and language varieties, with a large set of modalities (audio, videos, OCR, texts, lexica, etc.), and various annotations and tagging, etc. We also continue other identification tasks, either on projects like we do for the European Commission through the ELRC contract or through other internal initiatives. For instance, within the ELRC contract (`https://www.lr-coordination.eu/`), our consortium identified over 4300 resources, out of which 3450 are available to all users under permissive licenses e.g. CC and the like. I also mentioned that ELRA continues to identify and clear the IPR aspects for resources submitted to the LREMAP that inventories all resources listed by the authors of LREC papers since LREC 2010 and may other conferences that joined us. With almost 8000 items, many are available from their right holders or data centers. In addition to this, we continue to make available the set of Language Resources that have been trusted to us, often under permissive licenses, within the Share-Your-LRs initiative, an LREC feature.

## LR unique identification

A major problem that we are facing as a community is the duplicate resources in the different data centers that harvest online meta-data elements, producing their own inventories of data descriptions. In early 2000, a number of Language Resources players (inc. ELRA, LDC but also centers like Alaska Native Language Center, Langues et Civilisations à Tradition Orale, American Indian Studies Research Institute) agreed to join forces to establish OLAC, the Open Language Archives Community, that would be the repository of the meta-data to which researchers could refer.

With the emergence of a sizable number of infrastructures, many of them harvesting meta-data elements, it is essential that we assign unique identifiers to each LR to avoid that users get confused when acquiring or licensing resources.

Several major data centers have introduced the concept of ISLRN (International Standard Language Resource Number, https://islrn.org/), inspired from the publishing world. Since its inception, more than 200 institutions have joined this service, requesting ISLRN identifiers for their resources. We count now over 3200 identifiable resources in a unique way, wherever they are, even if they get renamed, etc. Of course ISLRN is part of the meta-data descriptors but it is more reliable than the naming and is not related to internet or to a given data-center or the internet-based Doi.

We hope to make this mandatory in the major conferences to ensure that we can keep track of existing resources that are distributed through multiple channels.

## Indigenous Languages

To strengthen its mission on processing all languages, ELRA continues to encourage high-quality publications on all language resources and on endangered ones at LREC as well as at the many satellite workshops or the Language Resources and Evaluation Journal (LRE journal). ELRA took an active part to the various debates that emerged a few years ago on Indigenous languages. The important one was initiated by the United Nations and its agency UNESCO, through the International Year of Indigenous Languages (IYIL), in 2019.

ELRA and its partners managed to bring together different communities with their respective expertise and experience related to language technologies for all at one of the most important events, the LT4ALL conference in 2019 (https://lt4all.org/en/). Through the organization of this first conference, we hope that we have open new doors and charted an important new path. We are very proud that the initial action has considerably grown to lead to the decade of Indigenous languages (2022-2032) and hopefully with strong commitments from public, civil society, private and government agencies, aiming to design an achievable roadmap for 2030. ELRA is very confident, now that the ELRA-ISCA joint SIG on the under-resourced languages (SIGUL) is fully operational and in good hands, to continue the work on this topic. More will be discussed during this conference and at the SIGUL workshop.

## LR production

ELDA has been and remains a major player in the production of language resources, both in R&D projects and for its partners. This production service covers all modalities, all languages, and all types of annotations. ELDA methodologies are designed with its partners from research and industry to ensure efficients production processes and high quality outcomes with a customized validation procedure. ELDA is proud to work with a large number of LREC participants to carry out such production activity.

## ELRA and Legal /Ethics

Part of the ELDA mission is to support the use of language resources or to conduct evaluation campaigns

and challenges in a clean legal context and conditions. Our experts continue to clear all IPR and other legal dimension to make the resources shareable. We continue to document these legal developments worldwide through regular reports but also through direct support to cases faced by the community (http://www.elra.info/en/dissemination/legal-issues-papers/). The main outcome of this is the very efficient partnership established with other data centers and experts to organize a useful workshop that takes place at each LREC. It is the community duty to ensure that all what is done complies with the legal frameworks but also adhere to high standards when it comes to ethics. ELRA, as a data center player, focuses its efforts on developing resources that reflect reality and accounting the ground truth that can help assess state of the art systems. But this ground truth should not compromise with the various biases that clearly impact the use of our technologies when based on data learning paradigms. The community needs to make our resources irreproachable, with transparent methodologies for the production as well as for the validation and the sharing. When the systems require a ground truth that clearly comprises bias, this should be explicitely documented, and an ideal situation contemplated.

This also applies to technology development. We need to ensure that systems' analysis and outputs are fair, leading to impeccably explainable results and compliant with our ethics.

Ethic is for many of us universal but likely, for many others, culturally based. Whatever we think, the use of our resources and technologies should aspire to a fundamental respect of humankind. This is why at each LREC, ELRA and its partners organize the specific workshop, mentioned above, to discuss ethical dimensions of our work, the novelties of legal frameworks all over the world, to brainstorm about requirements regarding respect of privacy, personal information, requirement for anonymization of our data sets, etc. LREC 2022 will continue this tradition through the dedicated legal & ethical issues workshop.

## Future Plans and Plans for the future

Where are we heading today? The new ELRA structure and governance will help continue to monitor the current developments. The community is heavily investing in IA-based packages and very large language models are produced and made available for various technologies and various languages even with multilingual approaches. ELRA supports such initiatives and is working to make available such Language Models, both the ones freely available under permissive licenses and the proprietary ones that could be license with particular conditions and restrictions. It's important to seek re-usability and avoid re-training again and again similar models, given the energy consumption needed. Of course such modeling should take place for new languages and new domains.

While Language Modeling is the current state of the art, new techniques and approaches will emerge sooner or later so as a community we should not abandon fundamental research on how to develop high-quality raw resources usable beyond today's state of affairs as well as the zero-resources approaches that may help many under resourced languages.

## Acknowledgments

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

- Google (Diamond)
- SADILAR (Silver)
- Vocapia (Silver)
- 3M (Bronze)
- Emvista (Bronze)
- Expert.ai (Bronze)
- Grammarly (Bronze)
- Délégation générale à la langue française et aux langues de France, Ministère de la Culture, France (Institutional Support, French Gouvernment)
- Région Sud Provence-Alpes-Côte d'Azur (Institutional Support, Local Gouvernment)
- MultiLingual (Media)

I would like also thank our colleagues from the local commitee, chaired by Frédéric Béchet and Philippe Blache. The local team was involved in the organization of two consecutive LRECs (2020 and 2022). 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. Thank you also to the ILC-CNR and ELDA teams who contributed to the scientific and organizational aspects of this LREC: Victoria Arranz, Roberto Bartolini, Lucille Blanchard, Francesca Frontini, Valérie Mapelli, Monica Monachini, Vincenzo Parrinelli, Valeria Quochi, Caroline Rannaud, Mickaël Rigaud, Alexandre Sicard, Kamelia Yahiaoui.

## Another word of thanks to Knut

A very particular gratefulness and gratitude on behalf of all of us, for the work carried out for more than thirty years. Knut Hofland modestly wrote *"I started the Corpora list after the 13th ICAME conference in Nijmegen in 1992 with 62 subscribers. The list has now more than 6200 subscribers".* Knut deserves our recognition for his excellent initiative, and all the efforts he did put in maintaining and moderating the Corpora list over these 30 years. ELRA is proud and prepared to take over, with the support of other institutions.

I hope that you will enjoy this 13th edition of the LREC conference, organized as a hybride event to allow more colleagues to join us, I hope that you will also appreciate the workshops that are going on site and virtual,

Marseille, France, June 21, 2022

Khalid Choukri

ELRA Secretary General and ELDA Chief Executive Officer

# LREC 2022

## 13th International Conference on Language Resources and Evaluation

## Opening Address

**Chairs of th Local Committee**
**Frédéric Béchet and Philippe Blache**

# LREC 2022 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
  - Stelios Piperidis  ILSP, Athens, Greece
- 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
  - Victoria Arranz  ELDA/ELRA, Paris, France
  - Roberto Bartolini  ILC/CNR, Pisa, Italy
  - Lucille Blanchard  ELDA/ELRA, Paris, France
  - Francesca Frontini  ILC/CNR, Pisa, Italy
  - Valérie Mapelli  ELDA/ELRA, Paris, France
  - Monica Monachini  ILC/CNR, Pisa, Italy
  - Vincenzo Parrinelli  ILC/CNR, Pisa, Italy
  - Valeria Quochi  ILC/CNR, Pisa, Italy
  - Caroline Rannaud  ELDA/ELRA, Paris, France
  - Mickaël Rigaud  ELDA/ELRA, Paris, France
  - Alexandre Sicard  ELDA/ELRA, Paris, France
  - Kamelia Yahiaoui  ELDA/ELRA, Paris, France
- Tutorial Chairs
  - Girish Jha  Jawaharlal Nehru University, India
  - Petya Osenova  Sofia University, Bulgaria
- Workshop Chairs
  - Simonetta Montemagni  ILC-CNR, Italy
  - Thamar Solorio  University of Houston, USA
  - Takenobu Tokunaga  Tokyo Institute of Technology, Japan

# LREC 2022 Area Chairs

- **Applications involving LRs and Evaluation (including applications in specific domains)**
  - Axel Ngonga, Paderborn University, Paderborn, Germany
  - Ivandré Paraboni, University of São Paulo, São Paulo, Brazil
  - Paul Rodrigues, Accenture, Washington (DC), USA & University of Maryland, College Park (MD), USA
  - Karin Verspoor, RMIT University, Melbourne, Australia

- **Corpora and Annotation (including Tools, Systems, Treebanks)**
  - Masayuki Asahara, National Institute for Japanese Language and Linguistics, Tolyo, Japan
  - Darja Fišer, University of Ljubljana and Institute of Contemporary History, Ljubljana, Slovenia
  - Monica Monachini, Istituto di Linguistica Computazionale "Antonio Zampolli", Pisa, Italy
  - Djamé Seddah, Université Paris-Sorbonne, Paris, France

- **Dialogue, Conversational Systems, Chatbots, Human-Robot Interaction**
  - Kristiina Jokinen, AI Research Center AIST / University of Helsinki, Tokyo, Japan
  - Sophie Rosset, Université Paris-Saclay, CNRS, LISN, Orsay, France
  - Jianhua Tao, Chinese Academy of Sciences, Beijing, China

- **Digital Humanities and Cultural Heritage**
  - Antske Fokkens, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
  - Sara Tonelli, Fondazione Bruno Kessler, Trento, Italy
  - Andreas Witt, Universität zu Köln, Köln, Germany

- **Discourse and Pragmatics**
  - Mikel Iruskieta, University of the Basque Country, Donostia, Spain
  - Leila Kosseim, Concordia University, Montreal, Canada
  - Deniz Zeyrek, Middle East Technical University, Graduate School of Informatics, Cognitive Science Department, Ankara, Turkey

- **Evaluation and Validation Methodologies**
  - Nicola Ferro, Department of Information Engineering - University of Padua, Padua, Italy
  - Constantin Orăsan, University of Surrey, Guildford, UK
  - Marcos Zampieri, Rochester Institute of Technology, Rochester (NY), USA

- **Information Extraction and Information Retrieval (including NER, QA, Text Mining, Document Classification, Text Categorisation)**
  - Wenliang Chen, Soochow University, Suzhou, China
  - Yakov Kronrod, Spotify, Florence (MA), USA
  - Paolo Rosso, Universitat Politècnica de València, València, Spain

- **Knowledge Discovery/Representation (including Knowledge Graphs, Linked Data, Terminology, Ontologies)**
  - Dagmar Gromann, University of Vienna, Vienna, Austria
  - German Rigau, HiTZ, Basque Center for Language Technology. University of the Basque Country, Donostia / San Sebastián, Spain
  - Han Xianpei, Institute of Software, Chinese Academy of Sciences, Beijing, China
- **Language Resources and Evaluation for Psycholinguistics, Cognitive Linguistics and Linguistic Theories**
  - Vera Demberg, Saarland University, Saarbrücken, Germany
  - Dimitrios Kokkinakis, Department of Swedish (Språkbanken Text) & the Centre for Ageing and Health (AgeCap), Gothenburg, Sweden
  - Dimitris N. Metaxas, Rutgers The State University of New Jersey, Piscataway (NJ), USA
- **Language Resource Infrastructures, Standards for LRs, Metadata, Policy issues, Ethics, Legal Issues**
  - Gilles Adda, LISN-CNRS, Orsay, Francer
  - Franciska de Jong, CLARIN ERIC, Utrecht, The Netherlands
  - Denise Di Persio, Linguistic Data Consortium, University of Pennsylvania, Philadelphia, USA
  - Georg Rehm, German Research Centre for Artificial Intelligence (DFKI), Berlin, Germany
- **Less Resourced/Endangered Languages**
  - Antti Arppe, University of Alberta, Edmonton, Canada
  - Kalika Bali, Microsoft Research Labs, Bangalore, India
  - Claudia Soria, Istituto di Linguistica Computazionale "Antonio Zampolli", Pisa, Italy
- **Lexicons (also WordNet, FrameNet, Multimodal and Sign Language lexicons, etc.)**
  - Simon Krek, Jožef Stefan Institute, Ljubljana, Slovenia
  - Bolette Sandford Pedersen, University of Copenhagen, Copenhagen, Denmark
  - Gilles Sérasset, Université Grenoble Alpes, Grenoble, France
- **Multilinguality and Machine Translation (including Speech-to Speech translation)**
  - Maite Melero, Barcelona Supercomputing Center, Barcelona, Spain
  - Toshiaki Nakazawa, The University of Tokyo, Tokyo, Japan
  - Maja Popovic, Dublin City University, Dublin, Ireland
- **Multimodality and Cross-modality (including Sign Languages, Vision and other modalities) and Multimedia**
  - Eleni Efthimiou, Athena RC, Institute for Language and Speech Processing (ILSP), Department of Embodied Interaction and Robotics, Athens, Greece
  - Costanza Navarretta, University of Copenhagen, Copenhagen, Denmark
  - Veronica Perez-Rosas, University of Michigan, Ann Arbor (MI) USA

- **Natural Language Generation (including Summarization)**
  - Naoaki Okazaki, Tokyo Institute of Technology, Tokyo, Japan
  - Horacio Saggion, Universitat Pompeu Fabra, Barcelona, Spain
  - Yaji Sripada, University of Aberdeen, Aberdeen, Scotland, UK
- **Opinion Mining, Sentiment Analysis, Emotion Recognition/Generation**
  - Rodrigo Agerri, HiTZ Center - Ixa, University of the Basque Country UPV/EHU, Donostia-San Sebastian, Spain
  - Viviana Patti, University of Turin, Turin, Italy
  - Theresa A. Wilson, Hanover College, Hanover (IN), USA
- **Parsing, Tagging, Grammar, Syntax, Morphology**
  - Tetsuji Nakagawa, Google Research, Tokyo, Japan
  - Xipeng Qiu, Fudan University, Shanghai, China
  - Francis Tyers, Indiana University, Bloomington (IN), USA
- **Reproduction of Research Results in Science and Technology of Language**
  - Anya Belz, ADAPT/DCU, Dublin, Ireland
  - Antonio Branco, University of Lisbon, Lisbon, Portugal
  - Piek Vossen, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
- **Semantics (including Distributional Semantics, Word Sense Disambiguation, Coreference, etc.)**
  - Eduardo Blanco, Arizona State University, Tempe (AZ), USA
  - Nancy Ide, Vassar College, Poughkeepsie (NY), USA & Brandeis University, Waltham (MA), USA
  - Roberto Navigli, Sapienza University of Rome, Rome, Italy
- **Social Media Processing**
  - Núria Bel, Universitat Pompeu Fabra, Barcelona, Spain
  - Kentaro Inui, Tohoku University / RIKEN, Sendai, Japan
  - Alexandre Pauchet, Normandie University, INSA Rouen, LITIS, Saint-Étienne-du-Rouvray, France
- **Speech Resources and Processing (including Phonetic Databases, Phonology, Prosody)**
  - Ahmed Ali, Qatar Computing Research Institute (QCRI), Doha, Qatar
  - Sanjeev Khudanpur, The Johns Hopkins University, Baltimore, Maryland, USA
  - Sakti Sakriani, Japan Advanced Institute of Science and Technology (JAIST), Nomi, Ishikawa, Japan

- **Statistical Methods and Machine Learning for Language Technologies (including Language Models)**
  - Felice Dell'Orletta, Istituto di Linguistica Computazionale "Antonio Zampolli", Pisa, Italy
  - Vincent Ng, University of Texas at Dallas, Richardson (TX), USA
  - Fabio Massimo Zanzotto, University of Rome Tor Vergata, Rome, Italy

## LREC 2022 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 2022 web site (https://lrec2022.lrec-conf.org/en/).

# 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 (Diamond)
  - SADILAR (Silver)
  - Vocapia (Silver)
  - 3M (Bronze)
  - Emvista (Bronze)
  - Expert.ai (Bronze)
  - Grammarly (Bronze)
  - Délégation générale à la langue française et aux langues de France, Ministère de la Culture, France (Institutional Support, French Gouvernment)
  - Région Sud Provence-Alpes-Côte d'Azur (Institutional Support, Local Gouvernment)
  - MultiLingual (Media)
- 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)

**ELRA is very grateful to their sponsors and supporters**

**Diamond Sponsor**

Google

**Silver Sponsors**

SADiLaR
South African Centre for Digital Language Resources

VOCAPIA research

**Bronze Sponsors**

3M Science. Applied to Life.™

emvista

expert.ai

grammarly

**MEDIA Partner**

MultiLingual

# Institutional Support

# Table of Contents

*Domain Adaptation in Neural Machine Translation using a Qualia-Enriched FrameNet*
Alexandre Diniz da Costa, Mateus Coutinho Marim, Ely Matos and Tiago Timponi Torrent . . . . . 1

*HOPE: A Task-Oriented and Human-Centric Evaluation Framework Using Professional Post-Editing Towards More Effective MT Evaluation*
Serge Gladkoff and Lifeng Han . . . . . 13

*Priming Ancient Korean Neural Machine Translation*
chanjun park, Seolhwa Lee, Jaehyung Seo, Hyeonseok Moon, Sugyeong Eo and Heuiseok Lim . 22

*GECO-MT: The Ghent Eye-tracking Corpus of Machine Translation*
Toon Colman, Margot Fonteyne, Joke Daems, Nicolas Dirix and Lieve Macken . . . . . 29

*Introducing Frege to Fillmore: A FrameNet Dataset that Captures both Sense and Reference*
Levi Remijnse, Piek Vossen, Antske Fokkens and Sam Titarsolej . . . . . 39

*Compiling a Suitable Level of Sense Granularity in a Lexicon for AI Purposes: The Open Source COR Lexicon*
Bolette Pedersen, Nathalie Carmen Hau Sørensen, Sanni Nimb, Ida Flørke, Sussi Olsen and Thomas Troelsgård . . . . . 51

*Sense and Sentiment*
Francis Bond and Merrick Choo . . . . . 61

*Enriching Linguistic Representation in the Cantonese Wordnet and Building the New Cantonese Wordnet Corpus*
Ut Seong Sio and Luís Morgado da Costa . . . . . 70

*ZAEBUC: An Annotated Arabic-English Bilingual Writer Corpus*
Nizar Habash and David Palfreyman . . . . . 79

*Turkish Universal Conceptual Cognitive Annotation*
Necva Bölücü and Burcu Can . . . . . 89

*Introducing the CURLICAT Corpora: Seven-language Domain Specific Annotated Corpora from Curated Sources*
Tamás Váradi, Bence Nyéki, Svetla Koeva, Marko Tadić, Vanja Štefanec, Maciej Ogrodniczuk, Bartłomiej Nitoń, Piotr Pęzik, Verginica Barbu Mititelu, Elena Irimia, Maria Mitrofan, Dan Tufiș, Radovan Garabík, Simon Krek and Andraž Repar . . . . . 100

*RU-ADEPT: Russian Anonymized Dataset with Eight Personality Traits*
C. Anton Rytting, Valerie Novak, James R. Hull, Victor M. Frank, Paul Rodrigues, Jarrett G. W. Lee and Laurel Miller-Sims . . . . . 109

*CoQAR: Question Rewriting on CoQA*
Quentin Brabant, Gwénolé Lecorvé and Lina M. Rojas Barahona . . . . . 119

*User Interest Modelling in Argumentative Dialogue Systems*
Annalena Aicher, Nadine Gerstenlauer, Wolfgang Minker and Stefan Ultes . . . . . 127

*Every time I fire a conversational designer, the performance of the dialogue system goes down*
Giancarlo Xompero, Michele Mastromattei, Samir Salman, Cristina Giannone, Andrea Favalli, Raniero Romagnoli and Fabio Massimo Zanzotto . . . . . 137

*An Empirical Study on the Overlapping Problem of Open-Domain Dialogue Datasets*
Yuqiao Wen, Guoqing Luo and Lili Mou . . . . . 146

*Language Technologies for the Creation of Multilingual Terminologies. Lessons Learned from the SSHOC Project*
Federica Gamba, Francesca Frontini, Daan Broeder and Monica Monachini . . . . . 154

*How to be FAIR when you CARE: The DGS Corpus as a Case Study of Open Science Resources for Minority Languages*
Marc Schulder and Thomas Hanke . . . . . 164

*Italian NLP for Everyone: Resources and Models from EVALITA to the European Language Grid*
Valerio Basile, Cristina Bosco, Michael Fell, Viviana Patti and Rossella Varvara . . . . . 174

*Cross-Lingual Link Discovery for Under-Resourced Languages*
Michael Rosner, Sina Ahmadi, Elena-Simona Apostol, Julia Bosque-Gil, Christian Chiarcos, Milan Dojchinovski, Katerina Gkirtzou, Jorge Gracia, Dagmar Gromann, Chaya Liebeskind, Giedrė Valūnaitė Oleškevičienė, Gilles Sérasset and Ciprian-Octavian Truică . . . . . 181

*Angry or Sad ? Emotion Annotation for Extremist Content Characterisation*
Valentina Dragos, Delphine Battistelli, Aline Etienne and Yolène Constable . . . . . 193

*Identification of Multiword Expressions in Tweets for Hate Speech Detection*
Nicolas Zampieri, Carlos Ramisch, Irina Illina and Dominique Fohr . . . . . 202

*Causal Investigation of Public Opinion during the COVID-19 Pandemic via Social Media Text*
Michael Jantscher and Roman Kern . . . . . 211

*Misspelling Semantics in Thai*
Pakawat Nakwijit and Matthew Purver . . . . . 227

*Automatic Detection of Stigmatizing Uses of Psychiatric Terms on Twitter*
Véronique MORICEAU, Farah Benamara and Abdelmoumene Boumadane . . . . . 237

*CoVERT: A Corpus of Fact-checked Biomedical COVID-19 Tweets*
Isabelle Mohr, Amelie Wührl and Roman Klinger . . . . . 244

*XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond*
Francesco Barbieri, Luis Espinosa Anke and Jose Camacho-Collados . . . . . 258

*'Am I the Bad One'? Predicting the Moral Judgement of the Crowd Using Pre–trained Language Models*
Areej Alhassan, Jinkai Zhang and Viktor Schlegel . . . . . 267

*Generating Questions from Wikidata Triples*
Kelvin Han, Thiago Castro Ferreira and Claire Gardent . . . . . 277

*Evaluating Transformer Language Models on Arithmetic Operations Using Number Decomposition*
Matteo Muffo, Aldo Cocco and Enrico Bertino . . . . . 291

*Evaluating the Effects of Embedding with Speaker Identity Information in Dialogue Summarization*
Yuji Naraki, Tetsuya Sakai and Yoshihiko Hayashi . . . . . 298

*Perceived Text Quality and Readability in Extractive and Abstractive Summaries*
Julius Monsen and Evelina Rennes . . . . . 305

*Learning to Prioritize: Precision-Driven Sentence Filtering for Long Text Summarization*
Alex Mei, Anisha Kabir, Rukmini Bapat, John Judge, Tony Sun and William Yang Wang . . . . . 313

*Automating Horizon Scanning in Future Studies*
Tatsuya Ishigaki, Suzuko Nishino, Sohei Washino, Hiroki Igarashi, Yukari Nagai, Yuichi Washida and Akihiko Murai . . . . . 319

*ViHealthBERT: Pre-trained Language Models for Vietnamese in Health Text Mining*
Nguyen Minh, Vu Hoang Tran, Vu Hoang, Huy Duc Ta, Trung Huu Bui and Steven Quoc Hung Truong . . . . . 328

*Privacy-Preserving Graph Convolutional Networks for Text Classification*
Timour Igamberdiev and Ivan Habernal . . . . . 338

*ArMATH: a Dataset for Solving Arabic Math Word Problems*
Reem Alghamdi, Zhenwen Liang and Xiangliang Zhang . . . . . 351

*KIMERA: Injecting Domain Knowledge into Vacant Transformer Heads*
Benjamin Winter, Alexei Figueroa Rosero, Alexander Löser, Felix Alexander Gers and Amy Siu 363

*Distilling the Knowledge of Romanian BERTs Using Multiple Teachers*
Andrei-Marius Avram, Darius Catrina, Dumitru-Clementin Cercel, Mihai Dascalu, Traian Rebedea, Vasile Pais and Dan Tufis . . . . . 374

*Personalized Filled-pause Generation with Group-wise Prediction Models*
Yuta Matsunaga, Takaaki Saeki, Shinnosuke Takamichi and Hiroshi Saruwatari . . . . . 385

*Transformer versus LSTM Language Models trained on Uncertain ASR Hypotheses in Limited Data Scenarios*
Imran Sheikh, Emmanuel Vincent and Irina Illina . . . . . 393

*Out of Thin Air: Is Zero-Shot Cross-Lingual Keyword Detection Better Than Unsupervised?*
Boshko Koloski, Senja Pollak, Blaž Škrlj and Matej Martinc . . . . . 400

*Evaluating Pretraining Strategies for Clinical BERT Models*
Anastasios Lamproudis, Aron Henriksson and Hercules Dalianis . . . . . 410

*KazNERD: Kazakh Named Entity Recognition Dataset*
Rustem Yeshpanov, Yerbolat Khassanov and Huseyin Atakan Varol . . . . . 417

*Mitigating Dataset Artifacts in Natural Language Inference Through Automatic Contextual Data Augmentation and Learning Optimization*
Michail Mersinias and Panagiotis Valvis . . . . . 427

*Kompetencer: Fine-grained Skill Classification in Danish Job Postings via Distant Supervision and Transfer Learning*
Mike Zhang, Kristian Nørgaard Jensen and Barbara Plank . . . . . 436

*Semantic Role Labelling for Dutch Law Texts*
Roos Bakker, Romy A.N. van Drie, Maaike de Boer, Robert van Doesburg and Tom van Engers448

*English Language Spelling Correction as an Information Retrieval Task Using Wikipedia Search Statistics*
Kyle Goslin and Markus Hofmann . . . . . 458

*CrudeOilNews: An Annotated Crude Oil News Corpus for Event Extraction*
Meisin Lee, Lay-Ki Soon, Eu Gene Siew and Ly Fie Sugianto . . . . . 465

*Claim Extraction and Law Matching for COVID-19-related Legislation*
Niklas Dehio, Malte Ostendorff and Georg Rehm . . . . . 480

*Constructing A Dataset of Support and Attack Relations in Legal Arguments in Court Judgements using Linguistic Rules*
Basit Ali, Sachin Pawar, Girish Palshikar and Rituraj Singh . . . . . 491

*KIND: an Italian Multi-Domain Dataset for Named Entity Recognition*
Teresa Paccosi and Alessio Palmero Aprosio . . . . . 501

*Russian Jeopardy! Data Set for Question-Answering Systems*
Elena Mikhalkova and Alexander A. Khlyupin . . . . . 508

*Know Better – A Clickbait Resolving Challenge*
Benjamin Hättasch and Carsten Binnig . . . . . 515

*Valet: Rule-Based Information Extraction for Rapid Deployment*
Dayne Freitag, John Cadigan, Robert Sasseen and Paul Kalmar . . . . . 524

*Negation Detection in Dutch Spoken Human-Computer Conversations*
Tom Sweers, Iris Hendrickx and Helmer Strik . . . . . 534

*Reflections on 30 Years of Language Resource Development and Sharing*
Christopher Cieri, Mark Liberman, Sunghye Cho, Stephanie Strassel, James Fiumara and Jonathan Wright . . . . . 543

*Language Resources to Support Language Diversity – the ELRA Achievements*
Valérie Mapelli, Victoria Arranz, Khalid Choukri and Hélène Mazo . . . . . 551

*Ethical Issues in Language Resources and Language Technology – Tentative Categorisation*
Pawel Kamocki and Andreas Witt . . . . . 559

*Do we Name the Languages we Study? The #BenderRule in LREC and ACL articles*
Fanny Ducel, Karën Fort, Gaël Lejeune and Yves Lepage . . . . . 564

*Aspect-Based Emotion Analysis and Multimodal Coreference: A Case Study of Customer Comments on Adidas Instagram Posts*
Luna De Bruyne, Akbar Karimi, Orphee De Clercq, Andrea Prati and Veronique Hoste . . . . . 574

*Multi-source Multi-domain Sentiment Analysis with BERT-based Models*
Gabriel Roccabruna, Steve Azzolin and Giuseppe Riccardi . . . . . 581

*NaijaSenti: A Nigerian Twitter Sentiment Corpus for Multilingual Sentiment Analysis*
Shamsuddeen Hassan Muhammad, David Adelani, Anuoluwapo Aremu and Idris Abdulmumin 590

*A (Psycho-)Linguistically Motivated Scheme for Annotating and Exploring Emotions in a Genre-Diverse Corpus*
Aline Etienne, Delphine Battistelli and Gwénolé Lecorvé . . . . . 603

*Integrating a Phrase Structure Corpus Grammar and a Lexical-Semantic Network: the HOLINET Knowledge Graph*
Jean-Philippe Prost . . . . . 613

*On the Impact of Temporal Representations on Metaphor Detection*
Giorgio Ottolina, Matteo Luigi Palmonari, Manuel Vimercati and Mehwish Alam . . . . . 623

*Analysis and Prediction of NLP Models via Task Embeddings*
Damien Sileo and Marie-Francine Moens . . . . . 633

*Cross-lingual and Cross-domain Transfer Learning for Automatic Term Extraction from Low Resource Data*
Amir Hazem, Merieme Bouhandi, Florian Boudin and Beatrice Daille . . . . . 648

*Few-Shot Learning for Argument Aspects of the Nuclear Energy Debate*
Lena Jurkschat, Gregor Wiedemann, Maximilian Heinrich, Mattes Ruckdeschel and Sunna Torge 663

*MuLVE, A Multi-Language Vocabulary Evaluation Data Set*
Anik Jacobsen, Salar Mohtaj and Sebastian Möller . . . . . 673

*PLOD: An Abbreviation Detection Dataset for Scientific Documents*
Leonardo Zilio, Hadeel Saadany, Prashant Sharma, Diptesh Kanojia and Constantin Orăsan . . . 680

*Potential Idiomatic Expression (PIE)-English: Corpus for Classes of Idioms*
Tosin Adewumi, Roshanak Vadoodi, Aparajita Tripathy, Konstantina Nikolaido, Foteini Liwicki and Marcus Liwicki . . . . . 689

*LeSpell - A Multi-Lingual Benchmark Corpus of Spelling Errors to Develop Spellchecking Methods for Learner Language*
Marie Bexte, Ronja Laarmann-Quante, Andrea Horbach and Torsten Zesch . . . . . 697

*Subjective Text Complexity Assessment for German*
Laura Seiffe, Fares Kallel, Sebastian Möller, Babak Naderi and Roland Roller . . . . . 707

*Querying Interaction Structure: Approaches to Overlap in Spoken Language Corpora*
Elena Frick, Thomas Schmidt and Henrike Helmer . . . . . 715

*DiaBiz – an Annotated Corpus of Polish Call Center Dialogs*
Piotr Pęzik, Gosia Krawentek, Sylwia Karasińska, Paweł Wilk, Paulina Rybińska, Anna Cichosz, Angelika Peljak-Łapińska, Mikołaj Deckert and Michał Adamczyk . . . . . 723

*LaVA – Latvian Language Learner corpus*
Roberts Darģis, Ilze Auziņa, Inga Kaija, Kristīne Levāne-Petrova and Kristīne Pokratniece . . . . 727

*The EuroPat Corpus: A Parallel Corpus of European Patent Data*
Kenneth Heafield, Elaine Farrow, Jelmer van der Linde, Gema Ramírez-Sánchez and Dion Wiggins 732

*"Beste Grüße, Maria Meyer" — Pseudonymization of Privacy-Sensitive Information in Emails*
Elisabeth Eder, Michael Wiegand, Ulrike Krieg-Holz and Udo Hahn . . . . . 741

*Criteria for the Annotation of Implicit Stereotypes*
Wolfgang Schmeisser-Nieto, Montserrat Nofre and Mariona Taulé . . . . . 753

*Common Phone: A Multilingual Dataset for Robust Acoustic Modelling*
Philipp Klumpp, Tomas Arias, Paula Andrea Pérez-Toro, Elmar Noeth and Juan Orozco-Arroyave 763

*Curras + Baladi: Towards a Levantine Corpus*
Karim Al-Haff, Mustafa Jarrar, Tymaa Hammouda and Fadi Zaraket . . . . . 769

*Annotation Study of Japanese Judgments on Tort for Legal Judgment Prediction with Rationales*
Hiroaki Yamada, Takenobu Tokunaga, Ryutaro Ohara, Keisuke Takeshita and Mihoko Sumida 779

*Placing M-Phasis on the Plurality of Hate: A Feature-Based Corpus of Hate Online*
Dana Ruiter, Liane Reiners, Ashwin Geet D'Sa, Thomas Kleinbauer, Dominique Fohr, Irina Illina, Dietrich Klakow, Christian Schemer and Angeliki Monnier . . . . . 791

*ParCorFull2.0: a Parallel Corpus Annotated with Full Coreference*
Ekaterina Lapshinova-Koltunski, Pedro Augusto Ferreira, Elina Lartaud and Christian Hardmeier 805

*A Multi-Party Dialogue Ressource in French*
Maria Boritchev and Maxime Amblard . . . . . 814

*Bicleaner AI: Bicleaner Goes Neural*
Jaume Zaragoza-Bernabeu, Gema Ramírez-Sánchez, Marta Bañón and Sergio Ortiz Rojas . . . . 824

*Semi-automatically Annotated Learner Corpus for Russian*
Anisia Katinskaia, Maria Lebedeva, Jue Hou and Roman Yangarber . . . . . 832

*UniMorph 4.0: Universal Morphology*
Khuyagbaatar Batsuren, Omer Goldman, Salam Khalifa, Nizar Habash, Witold Kieraś, Gábor Bella, Brian Leonard, Garrett Nicolai, Kyle Gorman, Yustinus Ghanggo Ate, Maria Ryskina, Sabrina Mielke, Elena Budianskaya, Charbel El-Khaissi, Tiago Pimentel, Michael Gasser, William Abbott Lane, Mohit Raj, Matt Coler, Jaime Rafael Montoya Samame, Delio Siticonatzi Camaiteri, Esaú Zumaeta Rojas, Didier López Francis, Arturo Oncevay, Juan López Bautista, Gema Celeste Silva Villegas, Lucas Torroba Hennigen, Adam Ek, David Guriel, Peter Dirix, Jean-Philippe Bernardy, Andrey Scherbakov, Aziyana Bayyr-ool, Antonios Anastasopoulos, Roberto Zariquiey, Karina Sheifer, Sofya Ganieva, Hilaria Cruz, Ritván Karahóǧa, Stella Markantonatou, George Pavlidis, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Candy Angulo, Jatayu Baxi, Andrew Krizhanovsky, Natalia Krizhanovskaya, Elizabeth Salesky, Clara Vania, Sardana Ivanova, Jennifer White, Rowan Hall Maudslay, Josef Valvoda, Ran Zmigrod, Paula Czarnowska, Irene Nikkarinen, Aelita Salchak, brijesh bhatt, Christopher Straughn, Zoey Liu, Jonathan North Washington, Yuval Pinter, Duygu Ataman, Marcin Wolinski, Totok Suhardijanto, Anna Yablonskaya, Niklas Stoehr, Hossep Dolatian, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Aryaman Arora, Richard J. Hatcher, Ritesh Kumar, Jeremiah Young, Daria Rodionova, Anastasia Yemelina, Taras Andrushko, Igor Marchenko, Polina Mashkovtseva, Alexandra Serova, Emily Prud'hommeaux, Maria Nepomniashchaya, fausto giunchiglia, Eleanor Chodroff, Mans Hulden, Miikka Silfverberg, Arya D. McCarthy, David Yarowsky, Ryan Cotterell, Reut Tsarfaty and Ekaterina Vylomova . . . . . 840

*Textinator: an Internationalized Tool for Annotation and Human Evaluation in Natural Language Processing and Generation*
Dmytro Kalpakchi and Johan Boye . . . . . 856

*CyberAgressionAdo-v1: a Dataset of Annotated Online Aggressions in French Collected through a Role-playing Game*
Anaïs Ollagnier, Elena Cabrio, Serena Villata and Catherine Blaya . . . . . 867

*Finnish Hate-Speech Detection on Social Media Using CNN and FinBERT*
Md Saroar Jahan, Mourad Oussalah and Nabil Arhab . . . . . 876

*Empirical Analysis of Noising Scheme based Synthetic Data Generation for Automatic Post-editing*
Hyeonseok Moon, chanjun park, Seolhwa Lee, Jaehyung Seo, Jungseob Lee, Sugyeong Eo and Heuiseok Lim . . . . . 883

*Domain Mismatch Doesn't Always Prevent Cross-lingual Transfer Learning*
Daniel Edmiston, Phillip Keung and Noah A. Smith . . . . . 892

*Cross-Lingual Knowledge Transfer for Clinical Phenotyping*
Jens-Michalis Papaioannou, Paul Grundmann, Betty van Aken, Athanasios Samaras, Ilias Kyparissidis, George Giannakoulas, Felix Gers and Alexander Loeser . . . . . 900

*The Multilingual Microblog Translation Corpus: Improving and Evaluating Translation of User-Generated Text*
Paul McNamee and Kevin Duh . . . . . 910

*Multilingual and Multimodal Learning for Brazilian Portuguese*
Júlia Sato, Helena Caseli and Lucia Specia . . . . . 919

*LibriS2S: A German-English Speech-to-Speech Translation Corpus*
Pedro Jeuris and Jan Niehues . . . . . 928

*A Linguistically Motivated Test Suite to Semi-Automatically Evaluate German–English Machine Translation Output*
Vivien Macketanz, Eleftherios Avramidis, Aljoscha Burchardt, He Wang, Renlong Ai, Shushen Manakhimova, Ursula Strohriegel, Sebastian Möller and Hans Uszkoreit . . . . . 936

*Cross-lingual Transfer of Monolingual Models*
Evangelia Gogoulou, Ariel Ekgren, Tim Isbister and Magnus Sahlgren . . . . . 948

*Dataset of Student Solutions to Algorithm and Data Structure Programming Assignments*
Fynn Petersen-Frey, Marcus Soll, Louis Kobras, Melf Johannsen, Peter Kling and Chris Biemann 956

*Language Patterns and Behaviour of the Peer Supporters in Multilingual Healthcare Conversational Forums*
Ishani Mondal, Kalika Bali, Mohit Jain, Monojit Choudhury, Jacki O'Neill, Millicent Ochieng, Kagnoya Awori and Keshet Ronen . . . . . 963

*Frame Shift Prediction*
Zheng Xin Yong, Patrick D. Watson, Tiago Timponi Torrent, Oliver Czulo and Collin Baker . . 976

*CLeLfPC: a Large Open Multi-Speaker Corpus of French Cued Speech*
Brigitte BIGI, Maryvonne Zimmermann and Carine André . . . . . 987

*Samrómur Children: An Icelandic Speech Corpus*
Carlos Daniel Hernandez Mena, David Erik Mollberg, Michal Borský and Jón Guðnason . . . . . 995

*The Norwegian Parliamentary Speech Corpus*
Per Erik Solberg and Pablo Ortiz . . . . . 1003

*A Speech Recognizer for Frisian/Dutch Council Meetings*
Martijn Bentum, Louis ten Bosch, Henk van den Heuvel, Simone Wills, Domenique van der Niet, Jelske Dijkstra and Hans Van de Velde . . . . . 1009

*Elderly Conversational Speech Corpus with Cognitive Impairment Test and Pilot Dementia Detection Experiment Using Acoustic Characteristics of Speech in Japanese Dialects*
Meiko Fukuda, Ryota Nishimura, Maina Umezawa, Kazumasa Yamamoto, Yurie Iribe and Norihide Kitaoka . . . . . 1016

*A Spoken Drug Prescription Dataset in French for Spoken Language Understanding*
Ali Can Kocabiyikoglu, François Portet, Prudence Gibert, Hervé Blanchon, Jean-Marc Babouchkine and Gaëtan Gavazzi . . . . . 1023

*Towards an Open-Source Dutch Speech Recognition System for the Healthcare Domain*
Cristian Tejedor-García, Berrie van der Molen, Henk van den Heuvel, Arjan van Hessen and Toine Pieters . . . . . 1032

*A Dataset for Speech Emotion Recognition in Greek Theatrical Plays*
Maria Moutti, Sofia Eleftheriou, Panagiotis Koromilas and Theodoros Giannakopoulos . . . . . 1040

*Audiobook Dialogues as Training Data for Conversational Style Synthetic Voices*
Liisi Piits, Hille Pajupuu, Heete Sahkai, Rene Altrov, Liis Ermus, Kairi Tamuri, Indrek Hein, Meelis Mihkla, Indrek Kiissel, Egert Männisalu, Kristjan Suluste and Jaan Pajupuu . . . . . 1047

*Using a Knowledge Base to Automatically Annotate Speech Corpora and to Identify Sociolinguistic Variation*
Yaru WU, Fabian Suchanek, Ioana Vasilescu, Lori Lamel and Martine Adda-Decker . . . . . 1054

*Phone Inventories and Recognition for Every Language*
Xinjian Li, Florian Metze, David R. Mortensen, Alan W Black and Shinji Watanabe . . . . . 1061

*Constructing Parallel Corpora from COVID-19 News using MediSys Metadata*
Dimitrios Roussis, Vassilis Papavassiliou, Sokratis Sofianopoulos, Prokopis Prokopidis and Stelios Piperidis . . . . . 1068

*A Distant Supervision Corpus for Extracting Biomedical Relationships Between Chemicals, Diseases and Genes*
Dongxu Zhang, Sunil Mohan, Michaela Torkar and Andrew McCallum . . . . . 1073

*DrugEHRQA: A Question Answering Dataset on Structured and Unstructured Electronic Health Records For Medicine Related Queries*
Jayetri Bardhan, Anthony Colas, Kirk Roberts and Daisy Zhe Wang . . . . . 1083

*Efficiently and Thoroughly Anonymizing a Transformer Language Model for Dutch Electronic Health Records: a Two-Step Method*
Stella Verkijk and Piek Vossen . . . . . 1098

*BERTrade: Using Contextual Embeddings to Parse Old French*
Loïc Grobol, Mathilde Regnault, Pedro Ortiz Suarez, Benoît Sagot, Laurent Romary and Benoit Crabbé . . . . . 1104

*Out-of-Domain Evaluation of Finnish Dependency Parsing*
Jenna Kanerva and Filip Ginter . . . . . 1114

*TArC: Tunisian Arabish Corpus, First complete release*
elisa gugliotta and Marco Dinarelli . . . . . 1125

*Towards Universal Segmentations: UniSegments 1.0*
Zdeněk Žabokrtský, Niyati Bafna, Jan Bodnár, Lukáš Kyjánek, Emil Svoboda, Magda Ševčíková and Jonáš Vidra . . . . . 1137

*TeDDi Sample: Text Data Diversity Sample for Language Comparison and Multilingual NLP*
Steven Moran, Christian Bentz, Ximena Gutierrez-Vasques, Olga Sozinova and Tanja Samardzic 1150

*Leveraging a Bilingual Dictionary to Learn Wolastoqey Word Representations*
Diego Bear and Paul Cook . . . . . 1159

*Unmasking the Myth of Effortless Big Data - Making an Open Source Multi-lingual Infrastructure and Building Language Resources from Scratch*
Linda Wiechetek, Katri Hiovain-Asikainen, Inga Lill Sigga Mikkelsen, Sjur Moshagen, Flammie Pirinen, Trond Trosterud and Børre Gaup . . . . . 1167

*Building and curating conversational corpora for diversity-aware language science and technology*
Andreas Liesenfeld and Mark Dingemanse . . . . . 1178

*EPIC UdS - Creation and Applications of a Simultaneous Interpreting Corpus*
Heike Przybyl, Ekaterina Lapshinova-Koltunski, Katrin Menzel, Stefan Fischer and Elke Teich1193

*Development of a Benchmark Corpus to Support Entity Recognition in Job Descriptions*
Thomas Green, Diana Maynard and Chenghua Lin . . . . . 1201

*CAMIO: A Corpus for OCR in Multiple Languages*
Michael Arrigo, Stephanie Strassel, Nolan King, Thao Tran and Lisa Mason . . . . . 1209

*FABRA: French Aggregator-Based Readability Assessment toolkit*
Rodrigo Wilkens, David Alfter, Xiaoou Wang, Alice Pintard, Anaïs Tack, Kevin P. Yancey and Thomas François . . . . . 1217

*Towards Building a Spoken Dialogue System for Argument Exploration*
Annalena Aicher, Nadine Gerstenlauer, Isabel Feustel, Wolfgang Minker and Stefan Ultes . . . 1234

*FreeTalky: Don't Be Afraid! Conversations Made Easier by a Humanoid Robot using Persona-based Dialogue*
chanjun park, Yoonna Jang, Seolhwa Lee, Sungjin Park and Heuiseok Lim . . . . . 1242

*Self-Contained Utterance Description Corpus for Japanese Dialog*
Yuta Hayashibe . . . . . 1249

*DialCrowd 2.0: A Quality-Focused Dialog System Crowdsourcing Toolkit*
Jessica Huynh, Ting-Rui Chiang, Jeffrey Bigham and Maxine Eskenazi . . . . . 1256

*A Brief Survey of Textual Dialogue Corpora*
Hugo Gonçalo Oliveira, Patrícia Ferreira, Daniel Martins, Catarina Silva and Ana Alves . . . . . 1264

*A Unified Approach to Entity-Centric Context Tracking in Social Conversations*
Ulrich Rückert, Srinivas Sunkara, Abhinav Rastogi, Sushant Prakash and Pranav Khaitan . . . . 1275

*A Unifying View On Task-oriented Dialogue Annotation*
Vojtěch Hudeček, leon-paul Schaub, Daniel Stancl, Patrick Paroubek and Ondřej Dušek . . . . . 1286

*A Multi-source Graph Representation of the Movie Domain for Recommendation Dialogues Analysis*
Antonio Origlia, Martina Di Bratto, Maria Di Maro and Sabrina Mennella . . . . . 1297

*SHARE: A Lexicon of Harmful Expressions by Spanish Speakers*
Flor Miriam Plaza-del-Arco, Ana Belén Parras Portillo, Pilar López Úbeda, Beatriz Gil and María-Teresa Martín-Valdivia . . . . . 1307

*Wiktextract: Wiktionary as Machine-Readable Structured Data*
Tatu Ylonen . . . . . 1317

*NyLLex: A Novel Resource of Swedish Words Annotated with Reading Proficiency Level*
Daniel Holmer and Evelina Rennes . . . . . 1326

*Making a Semantic Event-type Ontology Multilingual*
Zdenka Uresova, Karolina Zaczynska, Peter Bourgonje, Eva Fučíková, Georg Rehm and Jan Hajic 1332

*NomVallex: A Valency Lexicon of Czech Nouns and Adjectives*
Veronika Kolářová and Anna Vernerová . . . . . 1344

*TZOS: an Online Terminology Database Aimed at Working on Basque Academic Terminology Collaboratively*
Izaskun Aldezabal, Jose Mari Arriola and Arantxa Otegi . . . . . 1353

*Animacy Denoting German Nouns: Annotation and Classification*
Manfred Klenner and Anne Göhring . . . . . 1360

*x-enVENT: A Corpus of Event Descriptions with Experiencer-specific Emotion and Appraisal Annotations*
Enrica Troiano, Laura Ana Maria Oberlaender, Maximilian Wegge and Roman Klinger . . . . . 1365

*Polar Quantification of Actor Noun Phrases for German*
Anne Göhring and Manfred Klenner . . . . . 1376

*Czech Dataset for Cross-lingual Subjectivity Classification*
Pavel Přibáň and Josef Steinberger . . . . . 1381

*RED v2: Enhancing RED Dataset for Multi-Label Emotion Detection*
Alexandra Ciobotaru, Mihai Vlad Constantinescu, Liviu P. Dinu and Stefan Dumitrescu . . . . . 1392

*Fine-Grained Error Analysis and Fair Evaluation of Labeled Spans*
Katrin Ortmann . . . . . 1400

*Probing Pre-trained Auto-regressive Language Models for Named Entity Typing and Recognition*
Elena V. Epure and Romain Hennequin . . . . . 1408

*Frustratingly Easy Performance Improvements for Low-resource Setups: A Tale on BERT and Segment Embeddings*
Rob van der Goot, Max Müller-Eberstein and Barbara Plank . . . . . 1418

*The Subject Annotations of the Danish Parliament Corpus (2009-2017) - Evaluated with Automatic Multi-label Classification*
Costanza Navarretta and Dorte Haltrup Hansen . . . . . 1428

*A Systematic Study Reveals Unexpected Interactions in Pre-Trained Neural Machine Translation*
Ashleigh Richardson and Janet Wiles . . . . . 1437

*Holistic Evaluation of Automatic TimeML Annotators*
Mustafa Ocal, Adrian Perez, Antonela Radas and Mark Finlayson . . . . . 1444

*Measuring Uncertainty in Translation Quality Evaluation (TQE)*
Serge Gladkoff, Irina Sorokina, Lifeng Han and Alexandra Alekseeva . . . . . 1454

*Challenging the Transformer-based models with a Classical Arabic dataset: Quran and Hadith*
Shatha Altammami and Eric Atwell . . . . . 1462

*Question Modifiers in Visual Question Answering*
William Britton, Somdeb Sarkhel and Deepak Venugopal . . . . . 1472

*Multimodal Pipeline for Collection of Misinformation Data from Telegram*
Jose Sosa and Serge Sharoff . . . . . 1480

*Identifying Tension in Holocaust Survivors' Interview: Code-switching/Code-mixing as Cues*
Xinyuan Xia, Lu Xiao, Kun Yang and Yueyue Wang . . . . . 1490

*Fine-tuning vs From Scratch: Do Vision & Language Models Have Similar Capabilities on Out-of-Distribution Visual Question Answering?*
Kristian Nørgaard Jensen and Barbara Plank . . . . . 1496

*Multilingual Image Corpus – Towards a Multimodal and Multilingual Dataset*
Svetla Koeva, Ivelina Stoyanova and Jordan Kralev . . . . . 1509

*Sign Language Production With Avatar Layering: A Critical Use Case over Rare Words*
Jung-Ho Kim, Eui Jun Hwang, Sukmin Cho, Du Hui Lee and Jong Park . . . . . 1519

*The VoxWorld Platform for Multimodal Embodied Agents*
Nikhil Krishnaswamy, William Pickard, Brittany Cates, Nathaniel Blanchard and James Pustejovsky . . . . . 1529

*MemoSen: A Multimodal Dataset for Sentiment Analysis of Memes*
Eftekhar Hossain, Omar Sharif and Mohammed Moshiul Hoque . . . . . 1542

*RUSAVIC Corpus: Russian Audio-Visual Speech in Cars*
Denis Ivanko, Alexandr Axyonov, Dmitry Ryumin, Alexey Kashevnik and Alexey Karpov . . . 1555

*A First Corpus of AZee Discourse Expressions*
Camille Challant and Michael Filhol . . . . . 1560

*BERTHA: Video Captioning Evaluation Via Transfer-Learned Human Assessment*
Luis Lebron, Yvette Graham, Kevin McGuinness, Konstantinos Kouramas and Noel E. O'Connor 1566

*Abstract Meaning Representation for Gesture*
Richard Brutti, Lucia Donatelli, Kenneth Lai and James Pustejovsky . . . . . 1576

*The GINCO Training Dataset for Web Genre Identification of Documents Out in the Wild*
Taja Kuzman, Peter Rupnik and Nikola Ljubešić . . . . . 1584

*The Spoken Language Understanding MEDIA Benchmark Dataset in the Era of Deep Learning: data updates, training and evaluation tools*
Gaëlle Laperrière, Valentin Pelloin, Antoine Caubrière, salima mdhaffar, Nathalie Camelin, Sahar Ghannay, Bassam Jabaian and Yannick Estève . . . . . 1595

*BasqueGLUE: A Natural Language Understanding Benchmark for Basque*
Gorka Urbizu, Iñaki San Vicente, Xabier Saralegi, Rodrigo Agerri and Aitor Soroa . . . . . 1603

*Resources and Experiments on Sentiment Classification for Georgian*
Nicolas Stefanovitch, Jakub Piskorski and Sopho Kharazi . . . . . 1613

*CoFiF Plus: A French Financial Narrative Summarisation Corpus*
Nadhem ZMANDAR, Tobias Daudert, Sina Ahmadi, Mahmoud El-Haj and Paul Rayson . . . . 1622

*Generating Extended and Multilingual Summaries with Pre-trained Transformers*
Rémi Calizzano, Malte Ostendorff, Qian Ruan and Georg Rehm . . . . . 1640

*MUSS: Multilingual Unsupervised Sentence Simplification by Mining Paraphrases*
Louis Martin, Angela Fan, Éric de la Clergerie, Antoine Bordes and Benoît Sagot . . . . . 1651

*Towards Understanding Gender-Seniority Compound Bias in Natural Language Generation*
Samhita Honnavalli, Aesha Parekh, Lily Ou, Sophie Groenwold, Sharon Levy, Vicente Ordonez and William Yang Wang . . . . . 1665

*Combining ELECTRA and Adaptive Graph Encoding for Frame Identification*
Fabio Tamburini . . . . . 1671

*Polysemy in Spoken Conversations and Written Texts*
Aina Garí Soler, Matthieu Labeau and Chloé Clavel . . . . . 1680

*Cross-Level Semantic Similarity for Serbian Newswire Texts*
Vuk Batanović and Maja Miličević Petrović . . . . . 1691

*Universal Proposition Bank 2.0*
Ishan Jindal, Alexandre Rademaker, Michał Ulewicz, Ha Linh, Huyen Nguyen, Khoi-Nguyen Tran, Huaiyu Zhu and Yunyao Li . . . . . 1700

*The Copenhagen Corpus of Eye Tracking Recordings from Natural Reading of Danish Texts*
Nora Hollenstein, Maria Barrett and Marina Björnsdóttir . . . . . 1712

*The Brooklyn Multi-Interaction Corpus for Analyzing Variation in Entrainment Behavior*
Andreas Weise, Matthew McNeill and Rivka Levitan . . . . . 1721

*Pro-TEXT: an Annotated Corpus of Keystroke Logs*
Aleksandra Miletic, Christophe Benzitoun, Georgeta Cislaru and Santiago Herrera-Yanez . . . 1732

*Work Hard, Play Hard: Collecting Acceptability Annotations through a 3D Game*
Federico Bonetti, Elisa Leonardelli, Daniela Trotta, Raffaele Guarasci and Sara Tonelli . . . . . 1740

*DiHuTra: a Parallel Corpus to Analyse Differences between Human Translations*
Ekaterina Lapshinova-Koltunski, Maja Popović and Maarit Koponen . . . . . 1751

*Data Expansion Using WordNet-based Semantic Expansion and Word Disambiguation for Cyberbullying Detection*
Md Saroar Jahan, Djamila Romaissa Beddiar, Mourad Oussalah and Muhidin Mohamed . . . . . 1761

*ALIGNMEET: A Comprehensive Tool for Meeting Annotation, Alignment, and Evaluation*
Peter Polák, Muskaan Singh, Anna Nedoluzhko and Ondřej Bojar . . . . . 1771

*KSoF: The Kassel State of Fluency Dataset – A Therapy Centered Dataset of Stuttering*
Sebastian Bayerl, Alexander Wolff von Gudenberg, Florian Hönig, Elmar Noeth and Korbinian Riedhammer . . . . . 1780

*EZCAT: an Easy Conversation Annotation Tool*
Gaël Guibon, Luce Lefeuvre, Matthieu Labeau and Chloé Clavel . . . . . 1788

*Spoken Language Treebanks in Universal Dependencies: an Overview*
Kaja Dobrovoljc . . . . . 1798

*LeConTra: A Learner Corpus of English-to-Dutch News Translation*
Bram Vanroy and Lieve Macken . . . . . 1807

*Annotating Attribution in Czech News Server Articles*
Barbora Hladka, Jiří Mírovský, Matyáš Kopp and Václav Moravec . . . . . 1817

*Xposition: An Online Multilingual Database of Adpositional Semantics*
Luke Gessler, Nathan Schneider, Joseph C. Ledford and Austin Blodgett . . . . . 1824

*A Study in Contradiction: Data and Annotation for AIDA Focusing on Informational Conflict in Russia-Ukraine Relations*
Jennifer Tracey, Ann Bies, Jeremy Getman, Kira Griffitt and Stephanie Strassel . . . . . 1831

*Annotating Verbal Multiword Expressions in Arabic: Assessing the Validity of a Multilingual Annotation Procedure*
Najet Hadj Mohamed, Cherifa Ben Khelil, Agata Savary, Iskandar keskes, Jean-Yves Antoine and Lamia Hadrich-Belguith . . . . . 1839

*Annotation of Communicative Functions of Short Feedback Tokens in Switchboard*
Carol Figueroa, Adaeze Adigwe, Magalie Ochs and Gabriel Skantze . . . . . 1849

*A Dataset of Offensive Language in Kosovo Social Media*
Adem Ajvazi and Christian Hardmeier . . . . . 1860

*The Arabic Parallel Gender Corpus 2.0: Extensions and Analyses*
Bashar Alhafni, Nizar Habash and Houda Bouamor . . . . . 1870

*The Engage Corpus: A Social Media Dataset for Text-Based Recommender Systems*
Daniel Cheng, Kyle Yan, Phillip Keung and Noah A. Smith . . . . . 1885

*Annotating Arguments in a Corpus of Opinion Articles*
Gil Rocha, Luís Trigo, Henrique Lopes Cardoso, Rui Sousa-Silva, Paula Carvalho, Bruno Martins and Miguel Won . . . . . 1890

*German Parliamentary Corpus (GerParCor)*
Giuseppe Abrami, Mevlüt Bagci, Leon Hammerla and Alexander Mehler . . . . . 1900

*NerKor+Cars-OntoNotes++*
Attila Novák and Borbála Novák . . . . . 1907

*A Comparative Cross Language View On Acted Databases Portraying Basic Emotions Utilising Machine Learning*
Felix Burkhardt, Anabell Hacker, Uwe Reichel, Hagen Wierstorf, Florian Eyben and Björn Schuller 1917

*Nkululeko: A Tool For Rapid Speaker Characteristics Detection*
Felix Burkhardt, Johannes Wagner, Hagen Wierstorf, Florian Eyben and Björn Schuller ..... 1925

*Speech Aerodynamics Database, Tools and Visualisation*
Shi YU, Clara Ponchard, Roland Trouville, Sergio Hassid and Didier Demolin ..... 1933

*PATATRA and PATAFreq: two French databases for the documentation of within-speaker variability in speech*
Cécile Fougeron, Nicolas Audibert, cedric Gendrot, Estelle Chardenon and Louise Wohmann 1939

*The Makerere Radio Speech Corpus: A Luganda Radio Corpus for Automatic Speech Recognition*
Jonathan Mukiibi, Andrew Katumba, Joyce Nakatumba-Nabende, Ali Hussein and Joshua Meyer 1945

*Far-Field Speaker Recognition Benchmark Derived From The DiPCo Corpus*
Mickael Rouvier and Mohammad Mohammadamini ..... 1955

*Evaluating Sampling-based Filler Insertion with Spontaneous TTS*
Siyang Wang, joakim gustafson and Éva Székely ..... 1960

*BEA-Base: A Benchmark for ASR of Spontaneous Hungarian*
Peter Mihajlik, Andras Balog, Tekla Etelka Graczi, Anna Kohari, Balázs Tarján and Katalin Mady 1970

*SNuC: The Sheffield Numbers Spoken Language Corpus*
Emma Barker, Jon Barker, Robert Gaizauskas, Ning Ma and Monica Lestari Paramita ..... 1978

*The ManDi Corpus: A Spoken Corpus of Mandarin Regional Dialects*
Liang Zhao and Eleanor Chodroff ..... 1985

*The Speed-Vel Project: a Corpus of Acoustic and Aerodynamic Data to Measure Droplets Emission During Speech Interaction*
Francesca Carbone, Gilles Bouchet, Alain Ghio, Thierry Legou, Carine André, muriel lalain, Sabrina Kadri, Caterina Petrone, Federica Procino and Antoine Giovanni ..... 1991

*Towards Speech-only Opinion-level Sentiment Analysis*
Annalena Aicher, Alisa Gazizullina, Aleksei Gusev, Yuri Matveev and Wolfgang Minker .... 2000

*At the Intersection of NLP and Sustainable Development: Exploring the Impact of Demographic-Aware Text Representations in Modeling Value on a Corpus of Interviews*
Goya van Boven, Stephanie Hirmer and Costanza Conforti ..... 2007

*A Study on the Ambiguity in Human Annotation of German Oral History Interviews for Perceived Emotion Recognition and Sentiment Analysis*
Michael Gref, Nike Matthiesen, Sreenivasa Hikkal Venugopala, Shalaka Satheesh, Aswinkumar Vijayananth, Duc Bach Ha, Sven Behnke and Joachim Köhler ..... 2022

*Detecting Optimism in Tweets using Knowledge Distillation and Linguistic Analysis of Optimism*
Ștefan Cobeli, Ioan-Bogdan Iordache, Shweta Yadav, Cornelia Caragea, Liviu P. Dinu and Dragoș Iliescu ..... 2032

*Dataset and Baseline for Automatic Student Feedback Analysis*
Missaka Herath, Kushan Chamindu, Hashan Maduwantha and Surangika Ranathunga ..... 2042

*EENLP: Cross-lingual Eastern European NLP Index*
Alexey Tikhonov, Alex Malkhasov, Andrey Manoshin, George-Andrei Dima, Réka Cserháti, Md.Sadek Hossain Asif and Matt Sárdi . . . . . 2050

*Slovene SuperGLUE Benchmark: Translation and Evaluation*
Aleš Žagar and Marko Robnik-Šikonja . . . . . 2058

*Speech Resources in the Tamasheq Language*
Marcely Zanon Boito, Fethi Bougares, Florentin Barbier, Souhir Gahbiche, Loïc Barrault, Mickael Rouvier and Yannick Estève . . . . . 2066

*Aesop's fable "The North Wind and the Sun" Used as a Rosetta Stone to Extract and Map Spoken Words in Under-resourced Languages*
elena knyazeva, Philippe Boula de Mareüil and Frédéric Vernier . . . . . 2072

*Multilingual Open Text Release 1: Public Domain News in 44 Languages*
Chester Palen-Michel, June Kim and Constantine Lignos . . . . . 2080

*TweetTaglish: A Dataset for Investigating Tagalog-English Code-Switching*
Megan Herrera, Ankit Aich and Natalie Parde . . . . . 2090

*Jojajovai: A Parallel Guarani-Spanish Corpus for MT Benchmarking*
Luis Chiruzzo, Santiago Góngora, Aldo Alvarez, Gustavo Giménez-Lugo, Marvin Agüero-Torales and Yliana Rodríguez . . . . . 2098

*Assessing Multilinguality of Publicly Accessible Websites*
Rinalds Vīksna, Inguna Skadiņa, Raivis Skadiņš, Andrejs Vasiljevs and Roberts Rozis . . . . . 2108

*A Methodology for Building a Diachronic Dataset of Semantic Shifts and its Application to QC-FR-Diac-V1.0, a Free Reference for French*
David Kletz, Philippe Langlais, François Lareau and Patrick Drouin . . . . . 2117

*CRASS: A Novel Data Set and Benchmark to Test Counterfactual Reasoning of Large Language Models*
Jörg Frohberg and Frank Binder . . . . . 2126

*Evaluating Gender Bias in Speech Translation*
Marta R. Costa-jussà, Christine Basta and Gerard I. Gállego . . . . . 2141

*Design Choices in Crowdsourcing Discourse Relation Annotations: The Effect of Worker Selection and Training*
Merel Scholman, Valentina Pyatkin, Frances Yung, Ido Dagan, Reut Tsarfaty and Vera Demberg 2148

*TBD3: A Thresholding-Based Dynamic Depression Detection from Social Media for Low-Resource Users*
Hrishikesh Kulkarni, Sean MacAvaney, Nazli Goharian and Ophir Frieder . . . . . 2157

*SpecNFS: A Challenge Dataset Towards Extracting Formal Models from Natural Language Specifications*
Sayontan Ghosh, Amanpreet Singh, Alex Merenstein, Wei Su, Scott A. Smolka, Erez Zadok and Niranjan Balasubramanian . . . . . 2166

*Argument Similarity Assessment in German for Intelligent Tutoring: Crowdsourced Dataset and First Experiments*
Xiaoyu Bai and Manfred Stede . . . . . 2177

*Leveraging Pre-trained Language Models for Gender Debiasing*
Nishtha Jain, Declan Groves, Lucia Specia and Maja Popović . . . . . 2188

*Unsupervised Embeddings with Graph Auto-Encoders for Multi-domain and Multilingual Hate Speech Detection*
Gretel Liz De la Peña Sarracén and Paolo Rosso . . . . . 2196

*FQuAD2.0: French Question Answering and Learning When You Don't Know*
Quentin Heinrich, Gautier Viaud and Wacim Belblidia . . . . . 2205

*Large-Scale Hate Speech Detection with Cross-Domain Transfer*
Cagri Toraman, Furkan Şahinuç and Eyup Yilmaz . . . . . 2215

*GLoHBCD: A Naturalistic German Dataset for Language of Health Behaviour Change on Online Support Forums*
Selina Meyer and David Elsweiler . . . . . 2226

*Creating a Data Set of Abstractive Summaries of Turn-labeled Spoken Human-Computer Conversations*
Iris Hendrickx . . . . . 2236

*OpenEL: An Annotated Corpus for Entity Linking and Discourse in Open Domain Dialogue*
Wen Cui, Leanne Rolston, Marilyn Walker and Beth Ann Hockey . . . . . 2245

*Collecting Visually-Grounded Dialogue with A Game Of Sorts*
Bram Willemsen, Dmytro Kalpakchi and Gabriel Skantze . . . . . 2257

*CoRoSeOf - An Annotated Corpus of Romanian Sexist and Offensive Tweets*
Diana Constantina Hoefels, Çağrı Çöltekin and Irina Diana Mădroane . . . . . 2269

*ArMIS - The Arabic Misogyny and Sexism Corpus with Annotator Subjective Disagreements*
Dina Almanea and Massimo Poesio . . . . . 2282

*Annotating Interruption in Dyadic Human Interaction*
Liu YANG, Catherine ACHARD and Catherine PELACHAUD . . . . . 2292

*The Causal News Corpus: Annotating Causal Relations in Event Sentences from News*
Fiona Anting Tan, Ali Hürriyetoğlu, Tommaso Caselli, Nelleke Oostdijk, Tadashi Nomoto, Hansi Hettiarachchi, Iqra Ameer, Onur Uca, Farhana Ferdousi Liza and Tiancheng Hu . . . . . 2298

*Samrómur: Crowd-sourcing large amounts of data*
Staffan Hedström, David Erik Mollberg, Ragnheiður Þórhallsdóttir and Jón Guðnason . . . . . 2311

*An Annotated Corpus of Textual Explanations for Clinical Decision Support*
Roland Roller, Aljoscha Burchardt, Nils Feldhus, Laura Seiffe, Klemens Budde, Simon Ronicke and Bilgin Osmanodja . . . . . 2317

*LARD: Large-scale Artificial Disfluency Generation*
Tatiana Passali, Thanassis Mavropoulos, Grigorios Tsoumakas, Georgios Meditskos and Stefanos Vrochidis . . . . . 2327

*The CRECIL Corpus: a New Dataset for Extraction of Relations between Characters in Chinese Multi-party Dialogues*
Yuru Jiang, Yang Xu, Yuhang Zhan, Weikai He, Yilin Wang, Zixuan Xi, Meiyun Wang, Xinyu Li, Yu Li and Yanchao Yu . . . . . 2337

*The Bahrain Corpus: A Multi-genre Corpus of Bahraini Arabic*
Dana Abdulrahim, Go Inoue, Latifa Shamsan, Salam Khalifa and Nizar Habash . . . . . 2345

*A Universal Dependencies Treebank of Ancient Hebrew*
Daniel Swanson and Francis Tyers . . . . . 2353

*Hate Speech Dynamics Against African descent, Roma and LGBTQI Communities in Portugal*
Paula Carvalho, Bernardo Cunha, Raquel Santos, Fernando Batista and Ricardo Ribeiro . . . . . 2362

*Evolving Large Text Corpora: Four Versions of the Icelandic Gigaword Corpus*
Starkaður Barkarson, Steinþór Steingrímsson and Hildur Hafsteinsdóttir . . . . . 2371

*A Pragmatics-Centered Evaluation Framework for Natural Language Understanding*
Damien Sileo, Philippe Muller, Tim Van de Cruys and Camille Pradel . . . . . 2382

*Conversational Analysis of Daily Dialog Data using Polite Emotional Dialogue Acts*
Chandrakant Bothe and Stefan Wermter . . . . . 2395

*Inducing Discourse Marker Inventories from Lexical Knowledge Graphs*
Christian Chiarcos . . . . . 2401

*Story Trees: Representing Documents using Topological Persistence*
Pantea Haghighatkhah, Antske Fokkens, Pia Sommerauer, Bettina Speckmann and Kevin Verbeek 2413

*Extracting and Analysing Metaphors in Migration Media Discourse: towards a Metaphor Annotation Scheme*
Ana Zwitter Vitez, Mojca Brglez, Marko Robnik Šikonja, Tadej Škvorc, Andreja Vezovnik and Senja Pollak . . . . . 2430

*DDisCo: A Discourse Coherence Dataset for Danish*
Linea Flansmose Mikkelsen, Oliver Kinch, Anders Jess Pedersen and Ophélie Lacroix . . . . . 2440

*LPAttack: A Feasible Annotation Scheme for Capturing Logic Pattern of Attacks in Arguments*
Farjana Sultana Mim, Naoya Inoue, Shoichi Naito, Keshav Singh and Kentaro Inui . . . . . 2446

*BeSt: The Belief and Sentiment Corpus*
Jennifer Tracey, Owen Rambow, Claire Cardie, Adam Dalton, Hoa Trang Dang, Mona Diab, Bonnie Dorr, Louise Guthrie, Magdalena Markowska, Smaranda Muresan, Vinodkumar Prabhakaran, Samira Shaikh and Tomek Strzalkowski . . . . . 2460

*MOTIF: Contextualized Images for Complex Words to Improve Human Reading*
Xintong Wang, Florian Schneider, Özge Alacam, Prateek Chaudhury and Chris Biemann . . . . 2468

*Challenges with Sign Language Datasets for Sign Language Recognition and Translation*
Mirella De Sisto, Vincent Vandeghinste, Santiago Egea Gómez, Mathieu De Coster, Dimitar Shterionov and Horacio Saggion . . . . . 2478

*A Low-Cost Motion Capture Corpus in French Sign Language for Interpreting Iconicity and Spatial Referencing Mechanisms*
Clémence Mertz, Vincent BARREAUD, Thibaut Le Naour, Damien Lolive and Sylvie Gibet . 2488

*The CLAMS Platform at Work: Processing Audiovisual Data from the American Archive of Public Broadcasting*
Marc Verhagen, Kelley Lynch, Kyeongmin Rim and James Pustejovsky . . . . . 2498

*BU-NEmo: an Affective Dataset of Gun Violence News*
Carley Reardon, Sejin Paik, Ge Gao, Meet Parekh, Yanling Zhao, Lei Guo, Margrit Betke and Derry Tanti Wijaya . . . . . 2507

*RoomReader: A Multimodal Corpus of Online Multiparty Conversational Interactions*
Justine Reverdy, Sam O'Connor Russell, Louise Duquenne, Diego Garaialde, Benjamin R. Cowan and Naomi Harte . . . . . 2517

*Quevedo: Annotation and Processing of Graphical Languages*
Antonio F. G. Sevilla, Alberto Díaz Esteban and José María Lahoz-Bengoechea . . . . . 2528

*Merkel Podcast Corpus: A Multimodal Dataset Compiled from 16 Years of Angela Merkel's Weekly Video Podcasts*
Debjoy Saha, Shravan Nayak and Timo Baumann . . . . . 2536

*Crowdsourcing Kazakh-Russian Sign Language: FluentSigners-50*
Medet Mukushev, Aigerim Kydyrbekova, Alfarabi Imashev, Vadim Kimmelman and Anara Sandygulova 2541

*Connecting a French Dictionary from the Beginning of the 20th Century to Wikidata*
Pierre Nugues . . . . . 2548

*Metaphor annotation for German*
Markus Egg and Valia Kordoni . . . . . 2556

*NorDiaChange: Diachronic Semantic Change Dataset for Norwegian*
Andrey Kutuzov, Samia Touileb, Petter Mæhlum, Tita Enstad and Alexandra Wittemann . . . . 2563

*Exploring Transformers for Ranking Portuguese Semantic Relations*
Hugo Gonçalo Oliveira . . . . . 2573

*Building Static Embeddings from Contextual Ones: Is It Useful for Building Distributional Thesauri?*
Olivier Ferret . . . . . 2583

*Sentence Selection Strategies for Distilling Word Embeddings from BERT*
Yixiao Wang, Zied Bouraoui, Luis Espinosa Anke and Steven Schockaert . . . . . 2591

*DiaWUG: A Dataset for Diatopic Lexical Semantic Variation in Spanish*
Gioia Baldissin, Dominik Schlechtweg and Sabine Schulte im Walde . . . . . 2601

*My Case, For an Adposition: Lexical Polysemy of Adpositions and Case Markers in Finnish and Latin*
Daniel Chen and Mans Hulden . . . . . 2610

*WiC-TSV-de: German Word-in-Context Target-Sense-Verification Dataset and Cross-Lingual Transfer Analysis*
Anna Breit, Artem Revenko and Narayani Blaschke . . . . . 2617

*Re-train or Train from Scratch? Comparing Pre-training Strategies of BERT in the Medical Domain*
Hicham El Boukkouri, Olivier Ferret, Thomas Lavergne and Pierre Zweigenbaum . . . . . 2626

*Universal Semantic Annotator: the First Unified API for WSD, SRL and Semantic Parsing*
Riccardo Orlando, Simone Conia, Stefano Faralli and Roberto Navigli . . . . . 2634

*D3: A Massive Dataset of Scholarly Metadata for Analyzing the State of Computer Science Research*
Jan Philip Wahle, Terry Ruas, Saif Mohammad and Bela Gipp . . . . . 2642

*SciPar: A Collection of Parallel Corpora from Scientific Abstracts*
Dimitrios Roussis, Vassilis Papavassiliou, Prokopis Prokopidis, Stelios Piperidis and Vassilis Katsouros . . . . . 2652

*CATs are Fuzzy PETs: A Corpus and Analysis of Potentially Euphemistic Terms*
Martha Gavidia, Patrick Lee, Anna Feldman and JIng Peng . . . . . 2658

*Camel Treebank: An Open Multi-genre Arabic Dependency Treebank*
Nizar Habash, Muhammed AbuOdeh, Dima Taji, Reem Faraj, Jamila El Gizuli and Omar Kallas 2672

*MentSum: A Resource for Exploring Summarization of Mental Health Online Posts*
Sajad Sotudeh, Nazli Goharian and Zachary Young . . . . . 2682

*Klexikon: A German Dataset for Joint Summarization and Simplification*
Dennis Aumiller and Michael Gertz . . . . . 2693

*Applying Automatic Text Summarization for Fake News Detection*
Philipp Hartl and Udo Kruschwitz . . . . . 2702

*Increasing CMDI's Semantic Interoperability with schema.org*
Nino Meisinger, Thorsten Trippel and Claus Zinn . . . . . 2714

*RefCo and its Checker: Improving Language Documentation Corpora's Reusability Through a Semi-Automatic Review Process*
Herbert Lange and Jocelyn Aznar . . . . . 2721

*Identification and Analysis of Personification in Hungarian: The PerSECorp project*
Gábor Simon . . . . . 2730

*ISO-based Annotated Multilingual Parallel Corpus for Discourse Markers*
Purificação Silvano, Mariana Damova, Giedrė Valūnaitė Oleškevičienė, Chaya Liebeskind, Christian Chiarcos, Dimitar Trajanov, Ciprian-Octavian Truică, Elena-Simona Apostol and Anna Baczkowska 2739

*LIP-RTVE: An Audiovisual Database for Continuous Spanish in the Wild*
David Gimeno-Gómez and Carlos-D. Martínez-Hinarejos . . . . . 2750

*Modality Alignment between Deep Representations for Effective Video-and-Language Learning*
Hyeongu Yun, Yongil Kim and Kyomin Jung . . . . . 2759

*Mutual Gaze and Linguistic Repetition in a Multimodal Corpus*
Anais Murat, Maria Koutsombogera and Carl Vogel . . . . . 2771

*Multidimensional Coding of Multimodal Languaging in Multi-Party Settings*
Christophe Parisse, Marion Blondel, Stéphanie Caët, Claire Danet, Coralie Vincent and Aliyah Morgenstern . . . . . 2781

*Constructing a Lexical Resource of Russian Derivational Morphology*
Lukáš Kyjánek, Olga Lyashevskaya, Anna Nedoluzhko, Daniil Vodolazsky and Zdeněk Žabokrtský 2788

*Using Linguistic Typology to Enrich Multilingual Lexicons: the Case of Lexical Gaps in Kinship*
Temuulen Khishigsuren, Gábor Bella, Khuyagbaatar Batsuren, Abed Alhakim Freihat, Nandu Chandran Nair, Amarsanaa Ganbold, Hadi Khalilia, Yamini Chandrashekar and fausto giunchiglia 2798

*Towards Latvian WordNet*
Peteris Paikens, Mikus Grasmanis, Agute Klints, Ilze Lokmane, Lauma Pretkalniņa, Laura Rituma, Madara Stāde and Laine Strankale . . . . . 2808

*Building Sentiment Lexicons for Mainland Scandinavian Languages Using Machine Translation and Sentence Embeddings*
Peng Liu, Cristina Marco and Jon Atle Gulla . . . . . 2816

*A Thesaurus-based Sentiment Lexicon for Danish: The Danish Sentiment Lexicon*
Sanni Nimb, Sussi Olsen, Bolette Pedersen and Thomas Troelsgård . . . . . 2826

*IndoUKC: A Concept-Centered Indian Multilingual Lexical Resource*
Nandu Chandran Nair, Rajendran S. Velayuthan, Yamini Chandrashekar, Gábor Bella and fausto giunchiglia . . . . . 2833

*Korean Language Modeling via Syntactic Guide*
Hyeondey Kim, Seonhoon Kim, INHO KANG, Nojun Kwak and Pascale Fung . . . . . 2841

*A Whole-Person Function Dictionary for the Mobility, Self-Care and Domestic Life Domains: a Seedset Expansion Approach*
Ayah Zirikly, Bart Desmet, Julia Porcino, Jonathan Camacho Maldonado, Pei-Shu Ho, Rafael Jimenez Silva and Maryanne Sacco . . . . . 2850

*Placing multi-modal, and multi-lingual Data in the Humanities Domain on the Map: the Mythotopia Geo-tagged Corpus*
Voula Giouli, Anna Vacalopoulou, Nikolaos Sidiropoulos, Christina Flouda, Athanasios Doupas, Giorgos Giannopoulos, Nikos Bikakis, Vassilis Kaffes and Gregory Stainhaouer . . . . . 2856

*An Architecture of resolving a multiple link path in a standoff-style data format to enhance the mobility of language resources*
Kazushi Ohya . . . . . 2865

*A Corpus of German Citizen Contributions in Mobility Planning: Supporting Evaluation Through Multidimensional Classification*
Julia Romberg, Laura Mark and Tobias Escher . . . . . 2874

*Overlooked Data in Typological Databases: What Grambank Teaches Us About Gaps in Grammars*
Jakob Lesage, Hannah J. Haynie, Hedvig Skirgård, Tobias Weber and Alena Witzlack-Makarevich 2884

*Hong Kong: Longitudinal and Synchronic Characterisations of Protest News between 1998 and 2020*
Arya D. McCarthy and Giovanna Maria Dora Dore . . . . . 2891

*Nunc profana tractemus. Detecting Code-Switching in a Large Corpus of 16th Century Letters*
Martin Volk, Lukas Fischer, Patricia Scheurer, Bernard Silvan Schroffenegger, Raphael Schwitter, Phillip Ströbel and Benjamin Suter . . . . . 2901

*Quality and Efficiency of Manual Annotation: Pre-annotation Bias*
Marie Mikulová, Milan Straka, Jan Štěpánek, Barbora Štěpánková and Jan Hajic . . . . . 2909

*A Comprehensive Evaluation and Correction of the TimeBank Corpus*
Mustafa Ocal, Antonela Radas, Jared Hummer, Karine Megerdoomian and Mark Finlayson . . 2919

*Evaluating Multilingual Sentence Representation Models in a Real Case Scenario*
Rocco Tripodi, Rexhina Blloshmi and Simon Levis Sullam . . . . . 2928

*Validity, Agreement, Consensuality and Annotated Data Quality*
Anaëlle Baledent, Yann Mathet, Antoine Widlöcher, Christophe Couronne and Jean-Luc Manguin 2940

*Impact Analysis of the Use of Speech and Language Models Pretrained by Self-Supersivion for Spoken Language Understanding*
salima mdhaffar, Valentin Pelloin, Antoine Caubrière, Gaëlle Laperriere, Sahar Ghannay, Bassam Jabaian, Nathalie Camelin and Yannick Estève . . . . . 2949

*JGLUE: Japanese General Language Understanding Evaluation*
Kentaro Kurihara, Daisuke Kawahara and Tomohide Shibata . . . . . 2957

*Using the LARA Little Prince to compare human and TTS audio quality*
Elham Akhlaghi, Ingibjörg Iða Auðunardóttir, Anna Bączkowska, Branislav Bédi, Hakeem Beedar, Harald Berthelsen, Cathy Chua, Catia Cucchiarin, Hanieh Habibi, Ivana Horváthová, Junta Ikeda, Christèle Maizonniaux, Neasa Ní Chiaráin, Chadi Raheb, Manny Rayner, John Sloan, Nikos Tsourakis and Chunlin Yao . . . . . 2967

*Cyberbullying Classifiers are Sensitive to Model-Agnostic Perturbations*
Chris Emmery, Ákos Kádár, Grzegorz Chrupała and Walter Daelemans . . . . . 2976

*Constructing Distributions of Variation in Referring Expression Type from Corpora for Model Evaluation*
T. Mark Ellison and Fahime Same . . . . . 2989

*Knowledge Graph Question Answering Leaderboard: A Community Resource to Prevent a Replication Crisis*
Aleksandr Perevalov, Xi Yan, Liubov Kovriguina, Longquan Jiang, Andreas Both and Ricardo Usbeck . . . . . 2998

*Multi-Task Learning for Cross-Lingual Abstractive Summarization*
Sho Takase and Naoaki Okazaki . . . . . 3008

*How Much Context Span is Enough? Examining Context-Related Issues for Document-level MT*
Sheila Castilho . . . . . 3017

*TANDO: A Corpus for Document-level Machine Translation*
Harritxu Gete, Thierry Etchegoyhen, David Ponce, Gorka Labaka, Nora Aranberri, Ander Corral, Xabier Saralegi, Igor Ellakuria and Maite Martin . . . . . 3026

*Unsupervised Machine Translation in Real-World Scenarios*
Ona de Gibert Bonet, Iakes Goenaga, Jordi Armengol-Estapé, Olatz Perez-de-Viñaspre, Carla Parra Escartín, Marina Sanchez, Mārcis Pinnis, Gorka Labaka and Maite Melero . . . . . 3038

*COVID-19 Mythbusters in World Languages*
Mana Ashida, Jin-Dong Kim and Lee Seunghun . . . . . 3048

*On the Multilingual Capabilities of Very Large-Scale English Language Models*
Jordi Armengol-Estapé, Ona de Gibert Bonet and Maite Melero . . . . . 3056

*Evaluating Subtitle Segmentation for End-to-end Generation Systems*
Alina Karakanta, François Buet, Mauro Cettolo and François Yvon . . . . . 3069

*Using Semantic Role Labeling to Improve Neural Machine Translation*
Reinhard Rapp . . . . . 3079

*A Deep Transfer Learning Method for Cross-Lingual Natural Language Inference*
Dibyanayan Bandyopadhyay, Arkadipta De, Baban Gain, Tanik Saikh and Asif Ekbal . . . . . 3084

*Simple TICO-19: A Dataset for Joint Translation and Simplification of COVID-19 Texts*
Matthew Shardlow and Fernando Alva-Manchego . . . . . 3093

*Building Comparable Corpora for Assessing Multi-Word Term Alignment*
Omar Adjali, Emmanuel Morin and Pierre Zweigenbaum . . . . . 3103

*Mean Machine Translations: On Gender Bias in Icelandic Machine Translations*
Agnes Sólmundsdóttir, Dagbjört Guðmundsdóttir, Lilja Björk Stefánsdóttir and Anton Ingason 3113

*An Analysis of Dialogue Act Sequence Similarity Across Multiple Domains*
Ayesha Enayet and Gita Sukthankar . . . . . 3122

*Constructing a Culinary Interview Dialogue Corpus with Video Conferencing Tool*
Taro Okahisa, Ribeka Tanaka, Takashi Kodama, Yin Jou Huang and Sadao Kurohashi . . . . . 3131

*UgChDial: A Uyghur Chat-based Dialogue Corpus for Response Space Classification*
Zulipiye Yusupujiang and Jonathan Ginzburg . . . . . 3140

*A Speculative and Tentative Common Ground Handling for Efficient Composition of Uncertain Dialogue*
Saki Sudo, Kyoshiro Asano, Koh Mitsuda, Ryuichiro Higashinaka and Yugo Takeuchi . . . . . 3150

*BaSCo: An Annotated Basque-Spanish Code-Switching Corpus for Natural Language Understanding*
Maia Aguirre, Laura García-Sardiña, Manex Serras, Ariane Méndez and Jacobo López . . . . . 3158

*ProDial – An Annotated Proactive Dialogue Act Corpus for Conversational Assistants using Crowdsourcing*
Matthias Kraus, Nicolas Wagner and Wolfgang Minker . . . . . 3164

*ELITR Minuting Corpus: A Novel Dataset for Automatic Minuting from Multi-Party Meetings in English and Czech*
Anna Nedoluzhko, Muskaan Singh, Marie Hledíková, Tirthankar Ghosal and Ondřej Bojar . . 3174

*Extracting Age-Related Stereotypes from Social Media Texts*
Kathleen C. Fraser, Svetlana Kiritchenko and Isar Nejadgholi . . . . . 3183

*Borrowing or Codeswitching? Annotating for Finer-Grained Distinctions in Language Mixing*
Elena Alvarez-Mellado and Constantine Lignos . . . . . 3195

*Multi-Aspect Transfer Learning for Detecting Low Resource Mental Disorders on Social Media*
Ana Sabina Uban, Berta Chulvi and Paolo Rosso . . . . . 3202

*ArCovidVac: Analyzing Arabic Tweets About COVID-19 Vaccination*
Hamdy Mubarak, Sabit Hassan, Shammur Absar Chowdhury and Firoj Alam . . . . . 3220

*FACTOID: A New Dataset for Identifying Misinformation Spreaders and Political Bias*
Flora Sakketou, Joan Plepi, Riccardo Cervero, Henri Jacques Geiss, Paolo Rosso and Lucie Flek 3231

*Multitask Learning for Grapheme-to-Phoneme Conversion of Anglicisms in German Speech Recognition*
Julia Pritzen, Michael Gref, Dietlind Zühlke and Christoph Andreas Schmidt . . . . . 3242

*SDS-200: A Swiss German Speech to Standard German Text Corpus*
Michel Plüss, Manuela Hürlimann, Marc Cuny, Alla Stöckli, Nikolaos Kapotis, Julia Hartmann, Malgorzata Anna Ulasik, Christian Scheller, Yanick Schraner, Amit Jain, Jan Deriu, Mark Cieliebak and Manfred Vogel . . . . . 3250

*Extracting Linguistic Knowledge from Speech: A Study of Stop Realization in 5 Romance Languages*
Yaru WU, Mathilde Hutin, Ioana Vasilescu, Lori Lamel and Martine Adda-Decker . . . . . 3257

*Overlaps and Gender Analysis in the Context of Broadcast Media*
Martin Lebourdais, Marie Tahon, Antoine LAURENT, Sylvain Meignier and Anthony Larcher3264

*A Semi-Automatic Approach to Create Large Gender- and Age-Balanced Speaker Corpora: Usefulness of Speaker Diarization & Identification.*
Rémi Uro, David Doukhan, Albert Rilliard, Laetitia Larcher, Anissa-Claire Adgharouamane, Marie Tahon and Antoine Laurent . . . . . 3271

*DiscoGeM: A Crowdsourced Corpus of Genre-Mixed Implicit Discourse Relations*
Merel Scholman, Tianai Dong, Frances Yung and Vera Demberg . . . . . 3281

*QT30: A Corpus of Argument and Conflict in Broadcast Debate*
Annette Hautli-Janisz, Zlata Kikteva, Wassiliki Siskou, Kamila Gorska, Ray Becker and Chris Reed 3291

*Scaling up Discourse Quality Annotation for Political Science*
Neele Falk and Gabriella Lapesa . . . . . 3301

*Clarifying Implicit and Underspecified Phrases in Instructional Text*
Talita Anthonio, Anna Sauer and Michael Roth . . . . . 3319

*Multilingual Pragmaticon: Database of Discourse Formulae*
Anton Buzanov, Polina Bychkova, Arina Molchanova, Anna Postnikova and Daria Ryzhova . 3331

*Distant Reading in Digital Humanities: Case Study on the Serbian Part of the ELTeC Collection*
Ranka Stanković, Cvetana Krstev, Branislava Šandrih Todorović, Dusko Vitas, Mihailo Skoric and Milica Ikonić Nešić . . . . . 3337

*Exploring Text Recombination for Automatic Narrative Level Detection*
Nils Reiter, Judith Sieker, Svenja Guhr, Evelyn Gius and Sina Zarrieß . . . . . 3346

*Automatic Normalisation of Early Modern French*
Rachel Bawden, Jonathan Poinhos, Eleni Kogkitsidou, Philippe Gambette, Benoît Sagot and Simon Gabay . . . . . 3354

*From FreEM to D'AlemBERT: a Large Corpus and a Language Model for Early Modern French*
Simon Gabay, Pedro Ortiz Suarez, Alexandre BARTZ, Alix Chagué, Rachel Bawden, Philippe Gambette and Benoît Sagot . . . . . 3367

*Detecting Multiple Transitions in Literary Texts*
Nuette Heyns and Menno van Zaanen . . . . . 3375

*BasqueParl: A Bilingual Corpus of Basque Parliamentary Transcriptions*
Nayla Escribano, Jon Ander Gonzalez, Julen Orbegozo-Terradillos, Ainara Larrondo-Ureta, Simón Peña-Fernández, Olatz Perez-de-Viñaspre and Rodrigo Agerri . . . . . 3382

*GerEO: A Large-Scale Resource on the Syntactic Distribution of German Experiencer-Object Verbs*
Johanna M. Poppek, Simon Masloch and Tibor Kiss . . . . . 3391

*ACT2: A multi-disciplinary semi-structured dataset for importance and purpose classification of citations*
Suchetha Nambanoor Kunnath, Valentin Stauber, Ronin Wu, David Pride, Viktor Botev and Petr Knoth . . . . . 3398

*Quantification Annotation in ISO 24617-12, Second Draft*
Harry Bunt, Maxime Amblard, Johan Bos, Karën Fort, Bruno Guillaume, Philippe de Groote, Chuyuan Li, Pierre Ludmann, Michel Musiol, Siyana Pavlova, Guy Perrier and Sylvain Pogodalla 3407

*The LTRC Hindi-Telugu Parallel Corpus*
Vandan Mujadia and Dipti Sharma . . . . . 3417

*MHE: Code-Mixed Corpora for Similar Language Identification*
Priya Rani, John P. McCrae and Theodorus Fransen . . . . . 3425

*Bazinga! A Dataset for Multi-Party Dialogues Structuring*
Paul Lerner, Juliette Bergoënd, Camille Guinaudeau, Hervé Bredin, Benjamin Maurice, Sharleyne Lefevre, Martin Bouteiller, Aman Berhe, Léo Galmant, Ruiqing Yin and Claude Barras . . . . . 3434

*The Ellogon Web Annotation Tool: Annotating Moral Values and Arguments*
Alexandros Fotios Ntogramatzis, Anna Gradou, Georgios Petasis and Marko Kokol . . . . . 3442

*WeCanTalk: A New Multi-language, Multi-modal Resource for Speaker Recognition*
Karen Jones, Kevin Walker, Christopher Caruso, Jonathan Wright and Stephanie Strassel . . . . 3451

*Using Wiktionary to Create Specialized Lexical Resources and Datasets*
Lenka Bajčetić and Thierry Declerck . . . . . 3457

*STAPI: An Automatic Scraper for Extracting Iterative Title-Text Structure from Web Documents*
Nan Zhang, Shomir Wilson and Prasenjit Mitra . . . . . 3461

*ELTE Poetry Corpus: A Machine Annotated Database of Canonical Hungarian Poetry*
Péter Horváth, Péter Kundráth, Balázs Indig, Zsófia Fellegi, Eszter Szlávich, Tímea Borbála Bajzát, Zsófia Sárközi-Lindner, Bence Vida, Aslihan Karabulut, Mária Timári and Gábor Palkó . . . . . 3471

*HAWP: a Dataset for Hindi Arithmetic Word Problem Solving*
Harshita Sharma, Pruthwik Mishra and Dipti Sharma . . . . . 3479

*The Bulgarian Event Corpus: Overview and Initial NER Experiments*
Petya Osenova, Kiril Simov, Iva Marinova and Melania Berbatova . . . . . 3491

*A Corpus for Commonsense Inference in Story Cloze Test*
Bingsheng Yao, Ethan Joseph, Julian Lioanag and Mei Si . . . . . 3500

*Lessons Learned from GPT-SW3: Building the First Large-Scale Generative Language Model for Swedish*
Ariel Ekgren, Amaru Cuba Gyllensten, Evangelia Gogoulou, Alice Heiman, Severine Verlinden, Joey Öhman, Fredrik Carlsson and Magnus Sahlgren . . . . . 3509

*Constrained Language Models for Interactive Poem Generation*
Andrei Popescu-Belis, Àlex Atrio, Valentin Minder, Aris Xanthos, Gabriel Luthier, Simon Mattei and Antonio Rodriguez . . . . . 3519

*ELF22: A Context-based Counter Trolling Dataset to Combat Internet Trolls*
Huije Lee, Young Ju NA, Hoyun Song, Jisu Shin and Jong Park . . . . . 3530

*Generating Textual Explanations for Machine Learning Models Performance: A Table-to-Text Task*
Isaac Ampomah, James Burton, Amir Enshaei and Noura Al Moubayed . . . . . 3542

*Barch: an English Dataset of Bar Chart Summaries*
Iza Škrjanec, Muhammad Salman Edhi and Vera Demberg . . . . . 3552

*Effectiveness of Data Augmentation and Pretraining for Improving Neural Headline Generation in Low-Resource Settings*
Matej Martinc, Syrielle Montariol, Lidia Pivovarova and Elaine Zosa . . . . . 3561

*Effectiveness of French Language Models on Abstractive Dialogue Summarization Task*
Yongxin Zhou, François Portet and Fabien Ringeval . . . . . 3571

*ALEXSIS: A Dataset for Lexical Simplification in Spanish*
Daniel Ferrés and Horacio Saggion . . . . . 3582

*The IARPA BETTER Program Abstract Task Four New Semantically Annotated Corpora from IARPA's BETTER Program*
Timothy Mckinnon and Carl Rubino . . . . . 3595

*A Named Entity Recognition Corpus for Vietnamese Biomedical Texts to Support Tuberculosis Treatment*
Uyen Phan, Phuong N.V Nguyen and Nhung Nguyen . . . . . 3601

*RaFoLa: A Rationale-Annotated Corpus for Detecting Indicators of Forced Labour*
Erick Mendez Guzman, Viktor Schlegel and Riza Batista-Navarro . . . . . 3610

*Wojood: Nested Arabic Named Entity Corpus and Recognition using BERT*
Mustafa Jarrar, Mohammed Khalilia and Sana Ghanem . . . . . 3626

*Cross-lingual Approaches for the Detection of Adverse Drug Reactions in German from a Patient's Perspective*
Lisa Raithel, Philippe Thomas, Roland Roller, Oliver Sapina, Sebastian Möller and Pierre Zweigenbaum . . . . . 3637

*GGPONC 2.0 - The German Clinical Guideline Corpus for Oncology: Curation Workflow, Annotation Policy, Baseline NER Taggers*
Florian Borchert, Christina Lohr, Luise Modersohn, Jonas Witt, Thomas Langer, Markus Follmann, Matthias Gietzelt, Bert Arnrich, Udo Hahn and Matthieu-P. Schapranow . . . . . 3650

*ClinIDMap: Towards a Clinical IDs Mapping for Data Interoperability*
Elena Zotova, Montse Cuadros and German Rigau . . . . . 3661

*Identifying Draft Bills Impacting Existing Legislation: a Case Study on Romanian*
Corina Ceausu and Sergiu Nisioi . . . . . 3670

*MuLD: The Multitask Long Document Benchmark*
George Hudson and Noura Al Moubayed . . . . . 3675

*A Cross-document Coreference Dataset for Longitudinal Tracking across Radiology Reports*
Surabhi Datta, Hio Cheng Lam, Atieh Pajouhi, Sunitha Mogalla and Kirk Roberts . . . . . 3686

*How's Business Going Worldwide ? A Multilingual Annotated Corpus for Business Relation Extraction*
Hadjer Khaldi, Farah Benamara, Camille Pradel, Grégoire Sigel and Nathalie Aussenac-Gilles 3696

*Do Transformer Networks Improve the Discovery of Rules from Text?*
Mahdi Rahimi and Mihai Surdeanu . . . . . 3706

*Offensive language detection in Hebrew: can other languages help?*
Marina Litvak, Natalia Vanetik, Chaya Liebeskind, Omar Hmdia and Rizek Abu Madeghem . 3715

*JaMIE: A Pipeline Japanese Medical Information Extraction System with Novel Relation Annotation*
Fei Cheng, Shuntaro Yada, Ribeka Tanaka, Eiji ARAMAKI and Sadao Kurohashi . . . . . 3724

*Enhanced Entity Annotations for Multilingual Corpora*
Michael Strobl, Amine Trabelsi and Osmar Zaïane . . . . . 3732

*Enriching Epidemiological Thematic Features For Disease Surveillance Corpora Classification*
Edmond Menya, Mathieu Roche, Roberto Interdonato and Dickson Owuor . . . . . 3741

*Spanish Datasets for Sensitive Entity Detection in the Legal Domain*
Ona de Gibert Bonet, Aitor García Pablos, Montse Cuadros and Maite Melero . . . . . 3751

*ConvTextTM: An Explainable Convolutional Tsetlin Machine Framework for Text Classification*
Bimal Bhattarai, Ole-Christoffer Granmo and Lei Jiao . . . . . 3761

*Elvis vs. M. Jackson: Who has More Albums? Classification and Identification of Elements in Comparative Questions*
Meriem Beloucif, Seid Muhie Yimam, Steffen Stahlhacke and Chris Biemann . . . . . 3771

*Decorate the Examples: A Simple Method of Prompt Design for Biomedical Relation Extraction*
Hui-Syuan Yeh, Thomas Lavergne and Pierre Zweigenbaum . . . . . 3780

*Comparing Annotated Datasets for Named Entity Recognition in English Literature*
Rositsa Ivanova, Marieke van Erp and Sabrina Kirrane . . . . . 3788

*Investigating User Radicalization: A Novel Dataset for Identifying Fine-Grained Temporal Shifts in Opinion*
Flora Sakketou, Allison Lahnala, Liane Vogel and Lucie Flek . . . . . 3798

*APPReddit: a Corpus of Reddit Posts Annotated for Appraisal*
Marco Antonio Stranisci, Simona Frenda, Eleonora Ceccaldi, Valerio Basile, Rossana Damiano and Viviana Patti . . . . . 3809

*Evaluating Methods for Extraction of Aspect Terms in Opinion Texts in Portuguese - the Challenges of Implicit Aspects*
Mateus Machado and Thiago Alexandre Salgueiro Pardo . . . . . 3819

*SenticNet 7: A Commonsense-based Neurosymbolic AI Framework for Explainable Sentiment Analysis*
Erik Cambria, Qian Liu, Sergio Decherchi, Frank Xing and Kenneth Kwok . . . . . 3829

*Building an Endangered Language Resource in the Classroom: Universal Dependencies for Kakataibo*
Roberto Zariquiey, Claudia Alvarado, Ximena Echevarría, Luisa Gomez, Rosa Gonzales, Mariana Illescas, Sabina Oporto, Frederic Blum, Arturo Oncevay and Javier Vera . . . . . 3840

*The Norwegian Colossal Corpus: A Text Corpus for Training Large Norwegian Language Models*
Per Kummervold, Freddy Wetjen and Javier de la Rosa . . . . . 3852

*Embeddings models for Buddhist Sanskrit*
Ligeia Lugli, Matej Martinc, Andraž Pelicon and Senja Pollak . . . . . 3861

*Development of Automatic Speech Recognition for the Documentation of Cook Islands Māori*
Rolando Coto-Solano, Sally Akevai Nicholas, Samiha Datta, Victoria Quint, Piripi Wills, Emma Ngakuravaru Powell, Liam Koka'ua, Syed Tanveer and Isaac Feldman . . . . . 3872

*A Generalized Approach to Protest Event Detection in German Local News*
Gregor Wiedemann, Jan Matti Dollbaum, Sebastian Haunss, Priska Daphi and Larissa Daria Meier 3883

*Evaluation of Transfer Learning and Domain Adaptation for Analyzing German-Speaking Job Advertisements*
Ann-Sophie Gnehm, Eva Bühlmann and Simon Clematide . . . . . 3892

*Pre-Training Language Models for Identifying Patronizing and Condescending Language: An Analysis*
Carla Perez Almendros, Luis Espinosa Anke and Steven Schockaert . . . . . 3902

*HeLI-OTS, Off-the-shelf Language Identifier for Text*
Tommi Jauhiainen, Heidi Jauhiainen and Krister Lindén . . . . . 3912

*Towards a Broad Coverage Named Entity Resource: A Data-Efficient Approach for Many Diverse Languages*
Silvia Severini, Ayyoob ImaniGooghari, Philipp Dufter and Hinrich Schütze . . . . . 3923

*Towards the Construction of a WordNet for Old English*
Fahad Khan, Francisco J. Minaya Gómez, Rafael Cruz González, Harry Diakoff, Javier E. Diaz Vera, John P. McCrae, Ciara O'Loughlin, William Michael Short and Sander Stolk . . . . . 3934

*A Framenet and Frame Annotator for German Social Media*
Eckhard Bick . . . . . 3942

*The Robotic Surgery Procedural Framebank*
Marco Bombieri, Marco Rospocher, Simone Paolo Ponzetto and Paolo Fiorini . . . . . 3950

*Representing the Toddler Lexicon: Do the Corpus and Semantics Matter?*
Jennifer Weber and Eliana Colunga . . . . . 3960

*Organizing and Improving a Database of French Word Formation Using Formal Concept Analysis*
Nyoman Juniarta, Olivier Bonami, Nabil Hathout, Fiammetta Namer and Yannick Toussaint . 3969

*Towards a new Ontology for Sign Languages*
Thierry Declerck . . . . . 3977

*Towards the Detection of a Semantic Gap in the Chain of Commonsense Knowledge Triples*
Yoshihiko Hayashi . . . . . 3984

*COPA-SSE: Semi-structured Explanations for Commonsense Reasoning*
Ana Brassard, Benjamin Heinzerling, Pride Kavumba and Kentaro Inui . . . . . 3994

*GRhOOT: Ontology of Rhetorical Figures in German*
Ramona Kühn, Jelena Mitrović and Michael Granitzer . . . . . 4001

*Querying a Dozen Corpora and a Thousand Years with Fintan*
Christian Chiarcos, Christian Fäth and Maxim Ionov . . . . . 4011

*The Index Thomisticus Treebank as Linked Data in the LiLa Knowledge Base*
Francesco Mambrini, Marco Passarotti, Giovanni Moretti and Matteo Pellegrini . . . . . 4022

*Building a Multilingual Taxonomy of Olfactory Terms with Timestamps*
Stefano Menini, Teresa Paccosi, Serra Sinem Tekiroğlu and Sara Tonelli . . . . . 4030

*Attention Understands Semantic Relations*
Anastasia Chizhikova, Sanzhar Murzakhmetov, Oleg Serikov, Tatiana Shavrina and Mikhail Burtsev 4040

*Analysis of Dialogue in Human-Human Collaboration in Minecraft*
Takuma Ichikawa and Ryuichiro Higashinaka . . . . . 4051

*Data Collection for Empirically Determining the Necessary Information for Smooth Handover in Dialogue*
Sanae Yamashita and Ryuichiro Higashinaka . . . . . 4060

*The slurk Interaction Server Framework: Better Data for Better Dialog Models*
Jana Götze, Maike Paetzel-Prüsmann, Wencke Liermann, Tim Diekmann and David Schlangen 4069

*Corpus Design for Studying Linguistic Nudges in Human-Computer Spoken Interactions*
Natalia Kalashnikova, Serge Pajak, Fabrice Le Guel, Ioana Vasilescu, Gemma Serrano and Laurence Devillers . . . . . 4079

*Dialogue Corpus Construction Considering Modality and Social Relationships in Building Common Ground*
Yuki Furuya, Koki Saito, Kosuke Ogura, Koh Mitsuda, Ryuichiro Higashinaka and Kazunori Takashio . . . . . 4088

*EmoWOZ: A Large-Scale Corpus and Labelling Scheme for Emotion Recognition in Task-Oriented Dialogue Systems*
Shutong Feng, Nurul Lubis, Christian Geishauser, Hsien-chin Lin, Michael Heck, Carel van Niekerk and Milica Gasic . . . . . 4096

*Data Augmentation with Paraphrase Generation and Entity Extraction for Multimodal Dialogue System*
Eda Okur, Saurav Sahay and Lama Nachman . . . . . 4114

*Towards Modelling Self-imposed Filter Bubbles in Argumentative Dialogue Systems*
Annalena Aicher, Wolfgang Minker and Stefan Ultes . . . . . 4126

*Telling a Lie: Analyzing the Language of Information and Misinformation during Global Health Events*
Ankit Aich and Natalie Parde . . . . . 4135

*Misogyny and Aggressiveness Tend to Come Together and Together We Address Them*
Arianna Muti, Francesco Fernicola and Alberto Barrón-Cedeño . . . . . 4142

*The ComMA Dataset V0.2: Annotating Aggression and Bias in Multilingual Social Media Discourse*
Ritesh Kumar, Shyam Ratan, Siddharth Singh, Enakshi Nandi, Laishram Niranjana Devi, Akash Bhagat, Yogesh Dawer, bornini lahiri, Akanksha Bansal and Atul Kr. Ojha . . . . . 4149

*TUSC: Emotion Word Usage in Tweets from US and Canada*
Krishnapriya Vishnubhotla and Saif M. Mohammad . . . . . 4162

*A Turkish Hate Speech Dataset and Detection System*
Fatih Beyhan, Buse Çarık, İnanç Arın, Ayşecan Terzioğlu, Berrin Yanikoglu and Reyyan Yeniterzi 4177

*Life is not Always Depressing: Exploring the Happy Moments of People Diagnosed with Depression*
Ana-Maria Bucur, Adrian Cosma and Liviu P. Dinu . . . . . 4186

*Evaluating Tokenizers Impact on OOVs Representation with Transformers Models*
Alexandra Benamar, Cyril Grouin, Meryl Bothua and Anne Vilnat . . . . . 4193

*Assessing the Quality of an Italian Crowdsourced Idiom Corpus:the Dodiom Experiment*
Giuseppina Morza, Raffaele Manna and Johanna Monti . . . . . 4205

*Medical Crossing: a Cross-lingual Evaluation of Clinical Entity Linking*
Anton Alekseev, Zulfat Miftahutdinov, Elena Tutubalina, Artem Shelmanov, Vladimir Ivanov, Vladimir Kokh, Alexander Nesterov, Manvel Avetisian, Andrei Chertok and Sergey Nikolenko . . . 4212

*MTLens: Machine Translation Output Debugging*
Shreyas Sharma, Kareem Darwish, Lucas Pavanelli, Thiago Castro Ferreira, Mohamed Al-Badrashiny, Kamer Ali Yuksel and Hassan Sawaf . . . . . 4221

*IceBATS: An Icelandic Adaptation of the Bigger Analogy Test Set*
Steinunn Rut Friðriksdóttir, Hjalti Daníelsson, Steinþór Steingrímsson and Einar Sigurdsson . 4227

*Transfer Learning Methods for Domain Adaptation in Technical Logbook Datasets*
Farhad Akhbardeh, Marcos Zampieri, Cecilia Ovesdotter Alm and Travis Desell . . . . . 4235

*Downstream Task Performance of BERT Models Pre-Trained Using Automatically De-Identified Clinical Data*
Thomas Vakili, Anastasios Lamproudis, Aron Henriksson and Hercules Dalianis . . . . . 4245

*Dilated Convolutional Neural Networks for Lightweight Diacritics Restoration*
Bálint Csanády and András Lukács . . . . . 4253

*Generating Artificial Texts as Substitution or Complement of Training Data*
Vincent Claveau, Antoine Chaffin and Ewa Kijak . . . . . 4260

*From Pattern to Interpretation. Using Colibri Core to Detect Translation Patterns in the Peshitta.*
Mathias Coeckelbergs . . . . . 4270

*PAGnol: An Extra-Large French Generative Model*
Julien Launay, E.L. Tommasone, Baptiste Pannier, François Boniface, Amélie Chatelain, Alessandro Cappelli, Iacopo Poli and Djamé Seddah . . . . . 4275

*CEPOC: The Cambridge Exams Publishing Open Cloze dataset*
Mariano Felice, Shiva Taslimipoor, Øistein E. Andersen and Paula Buttery . . . . . 4285

*ALBETO and DistilBETO: Lightweight Spanish Language Models*
José Cañete, Sebastian Donoso, Felipe Bravo-Marquez, Andrés Carvallo and Vladimir Araujo 4291

*On the Robustness of Cognate Generation Models*
Winston Wu and David Yarowsky . . . . . 4299

*CLISTER : A Corpus for Semantic Textual Similarity in French Clinical Narratives*
Nicolas Hiebel, Olivier Ferret, Karën Fort and Aurélie Névéol . . . . . 4306

*The Chinese Causative-Passive Homonymy Disambiguation: an adversarial Dataset for NLI and a Probing Task*
Shanshan Xu and Katja Markert . . . . . 4316

*Modeling Noise in Paraphrase Detection*
Teemu Vahtola, Eetu Sjöblom, Jörg Tiedemann and Mathias Creutz . . . . . 4324

*Give me your Intentions, I'll Predict our Actions: A Two-level Classification of Speech Acts for Crisis Management in Social Media*
Enzo laurenti, Nils Bourgon, Farah Benamara, Alda Mari, Véronique MORICEAU and Camille Courgeon . . . . . 4333

*Towards a Cleaner Document-Oriented Multilingual Crawled Corpus*
Julien Abadji, Pedro Ortiz Suarez, Laurent Romary and Benoît Sagot . . . . . 4344

*A Warm Start and a Clean Crawled Corpus - A Recipe for Good Language Models*
Vésteinn Snæbjarnarson, Haukur Barri Símonarson, Pétur Orri Ragnarsson, Svanhvít Lilja Ingólfsdóttir, Haukur Jónsson, Vilhjalmur Thorsteinsson and Hafsteinn Einarsson . . . . . 4356

*Adapting Language Models When Training on Privacy-Transformed Data*
Tugtekin Turan, Dietrich Klakow, Emmanuel Vincent and Denis Jouvet . . . . . 4367

*Evaluation of Transfer Learning for Polish with a Text-to-Text Model*
Aleksandra Chrabrowa, Łukasz Dragan, Karol Grzegorczyk, Dariusz Kajtoch, Mikołaj Koszowski, Robert Mroczkowski and Piotr Rybak . . . . . 4374

*Evaluation of HTR models without Ground Truth Material*
Phillip Benjamin Ströbel, Martin Volk, Simon Clematide, Raphael Schwitter, Tobias Hodel and David Schoch . . . . . 4395

*A Semi-Automated Live Interlingual Communication Workflow Featuring Intralingual Respeaking: Evaluation and Benchmarking*
Tomasz Korybski, Elena Davitti, Constantin Orasan and Sabine Braun . . . . . 4405

*Are Embedding Spaces Interpretable? Results of an Intrusion Detection Evaluation on a Large French Corpus*
Thibault Prouteau, Nicolas Dugué, Nathalie Camelin and Sylvain Meignier . . . . . 4414

*Corpus for Automatic Structuring of Legal Documents*
Prathamesh Kalamkar, Aman Tiwari, Astha Agarwal, Saurabh Karn, Smita Gupta, Vivek Raghavan and Ashutosh Modi . . . . . 4420

*The Search for Agreement on Logical Fallacy Annotation of an Infodemic*
Claire Bonial, Austin Blodgett, Taylor Hudson, Stephanie M. Lukin, Jeffrey Micher, Douglas Summers-Stay, Peter Sutor and Clare Voss . . . . . 4430

*Recovering Patient Journeys: A Corpus of Biomedical Entities and Relations on Twitter (BEAR)*
Amelie Wührl and Roman Klinger . . . . . 4439

*Improving Event Duration Question Answering by Leveraging Existing Temporal Information Extraction Data*
Felix Virgo, Fei Cheng and Sadao Kurohashi . . . . . 4451

*Entity Linking over Nested Named Entities for Russian*
Natalia Loukachevitch, Pavel Braslavski, Vladimir Ivanov, Tatiana Batura, Suresh Manandhar, Artem Shelmanov and Elena Tutubalina . . . . . 4458

*HiNER: A large Hindi Named Entity Recognition Dataset*
Rudra Murthy, Pallab Bhattacharjee, Rahul Sharnagat, Jyotsana Khatri, Diptesh Kanojia and Pushpak Bhattacharyya . . . . . 4467

*Bootstrapping Text Anonymization Models with Distant Supervision*
Anthi Papadopoulou, Pierre Lison, Lilja Øvrelid and Ildikó Pilán . . . . . 4477

*Natural Questions in Icelandic*
Vésteinn Snæbjarnarson and Hafsteinn Einarsson . . . . . 4488

*QA4IE: A Quality Assurance Tool for Information Extraction*
Rafael Jimenez Silva, Kaushik Gedela, Alex Marr, Bart Desmet, Carolyn Rose and Chunxiao Zhou 4497

*A New Dataset for Topic-Based Paragraph Classification in Genocide-Related Court Transcripts*
Miriam Schirmer, Udo Kruschwitz and Gregor Donabauer . . . . . 4504

*DeepREF: A Framework for Optimized Deep Learning-based Relation Classification*
Igor Nascimento, Rinaldo Lima, Adrian-Gabriel CHIFU, Bernard Espinasse and Sébastien Fournier 4513

*Exploring Data Augmentation Strategies for Hate Speech Detection in Roman Urdu*
Ubaid Azam, Hammad Rizwan and Asim Karim . . . . . 4523

*Incorporating LIWC in Neural Networks to Improve Human Trait and Behavior Analysis in Low Resource Scenarios*
Isil Yakut Kilic and Shimei Pan . . . . . 4532

*Using Sentence-level Classification Helps Entity Extraction from Material Science Literature*
Ankan Mullick, Shubhraneel Pal, Tapas Nayak, Seung-Cheol Lee, Satadeep Bhattacharjee and Pawan Goyal . . . . . 4540

*A Twitter Corpus for Named Entity Recognition in Turkish*
Buse Çarık and Reyyan Yeniterzi . . . . . 4546

*A STEP towards Interpretable Multi-Hop Reasoning:Bridge Phrase Identification and Query Expansion*
Fan Luo and Mihai Surdeanu . . . . . 4552

*Question Generation and Answering for exploring Digital Humanities collections*
Frederic Bechet, Elie Antoine, Jérémy Auguste and Géraldine Damnati . . . . . 4561

*Evaluating Retrieval for Multi-domain Scientific Publications*
Nancy Ide, Keith Suderman, Jingxuan Tu, Marc Verhagen, Shanan Peters, Ian Ross, John Lawson, Andrew Borg and James Pustejovsky . . . . . 4569

*Modeling Dutch Medical Texts for Detecting Functional Categories and Levels of COVID-19 Patients*
Jenia Kim, Stella Verkijk, Edwin Geleijn, Marieke van der Leeden, Carel Meskers, Caroline Meskers, Sabina van der Veen, Piek Vossen and Guy Widdershoven . . . . . 4577

*Hierarchical Aggregation of Dialectal Data for Arabic Dialect Identification*
Nurpeiis Baimukan, Houda Bouamor and Nizar Habash . . . . . 4586

*Investigating Active Learning Sampling Strategies for Extreme Multi Label Text Classification*
Lukas Wertz, Katsiaryna Mirylenka, Jonas Kuhn and Jasmina Bogojeska . . . . . 4597

*German Light Verb Constructions in Business Process Models*
Kristin Kutzner and Ralf Laue . . . . . 4606

*PhysNLU: A Language Resource for Evaluating Natural Language Understanding and Explanation Coherence in Physics*
Jordan Meadows, Zili Zhou and André Freitas . . . . . 4611

*HECTOR: A Hybrid TExt SimplifiCation TOol for Raw Texts in French*
Amalia Todirascu, Rodrigo Wilkens, Eva Rolin, Thomas François, Delphine Bernhard and Núria Gala . . . . . 4620

*AiRO - an Interactive Learning Tool for Children at Risk of Dyslexia*
Peter Juel Henrichsen and Stine Fuglsang Engmose . . . . . 4631

*Creating a Basic Language Resource Kit for Faroese*
Annika Simonsen, Sandra Saxov Lamhauge, Iben Nyholm Debess and Peter Juel Henrichsen 4637

*Developing a Spell and Grammar Checker for Icelandic using an Error Corpus*
Hulda Óladóttir, Þórunn Arnardóttir, Anton Ingason and Vilhjálmur Þorsteinsson . . . . . 4644

*The TalkMoves Dataset: K-12 Mathematics Lesson Transcripts Annotated for Teacher and Student Discursive Moves*
Abhijit Suresh, Jennifer Jacobs, Charis Harty, Margaret Perkoff, James H. Martin and Tamara Sumner . . . . . 4654

*Automating Idea Unit Segmentation and Alignment for Assessing Reading Comprehension via Summary Protocol Analysis*
Marcello Gecchele, Hiroaki Yamada, Takenobu Tokunaga, Yasuyo Sawaki and Mika Ishizuka4663

*IRAC: A Domain-Specific Annotated Corpus of Implicit Reasoning in Arguments*
Keshav Singh, Naoya Inoue, Farjana Sultana Mim, Shoichi Naito and Kentaro Inui . . . . . 4674

*Conversational Speech Recognition Needs Data? Experiments with Austrian German*
Julian Linke, Philip N. Garner, Gernot Kubin and Barbara Schuppler . . . . . 4684

*A Benchmark Corpus for the Detection of Automatically Generated Text in Academic Publications*
Vijini Liyanage, Davide Buscaldi and Adeline Nazarenko . . . . . 4692

*Building a Dataset for Automatically Learning to Detect Questions Requiring Clarification*
Ivano Lauriola, Kevin Small and Alessandro Moschitti . . . . . 4701

*The ALPIN Sentiment Dictionary: Austrian Language Polarity in Newspapers*
Thomas Kolb, Sekanina Katharina, Bettina Manuela Johanna Kern, Julia Neidhardt, Tanja Wissik and Andreas Baumann . . . . . 4708

*Text Classification and Prediction in the Legal Domain*
Minh-Quoc Nghiem, Paul Baylis, André Freitas and Sophia Ananiadou . . . . . 4717

*I still have Time(s): Extending HeidelTime for German Texts*
Andy Luecking, Manuel Stoeckel, Giuseppe Abrami and Alexander Mehler . . . . . 4723

*Morphological Complexity of Children Narratives in Eight Languages*
Gordana Hržica, Chaya Liebeskind, Kristina Š. Despot, Olga Dontcheva-Navratilova, Laura Kamandulytė-Merfeldienė, Sara Košutar, Matea Kramarić and Giedrė Valūnaitė Oleškevičienė . . . . . 4729

*EXPRES Corpus for A Field-specific Automated Exploratory Study of L2 English Expert Scientific Writing*
Ana-Maria Bucur, Madalina Chitez, Valentina Muresan, Andreea Dinca and Roxana Rogobete4739

*An Evaluation Framework for Legal Document Summarization*
Ankan Mullick, Abhilash Nandy, Manav Kapadnis, Sohan Patnaik, Raghav R and Roshni Kar4747

*Complex Labelling and Similarity Prediction in Legal Texts: Automatic Analysis of France's Court of Cassation Rulings*
Thibault Charmet, Inès Cherichi, Matthieu Allain, Urszula Czerwinska, Amaury Fouret, Benoît Sagot and Rachel Bawden . . . . . 4754

*Cyrillic-MNIST: a Cyrillic Version of the MNIST Dataset*
Bolat Tleubayev, Zhanel Zhexenova, Kenessary Koishybay and Anara Sandygulova . . . . . 4767

*gaBERT — an Irish Language Model*
James Barry, Joachim Wagner, Lauren Cassidy, Alan Cowap, Teresa Lynn, Abigail Walsh, Mícheál J. Ó Meachair and Jennifer Foster . . . . . 4774

*PoS Tagging, Lemmatization and Dependency Parsing of West Frisian*
Wilbert Heeringa, Gosse Bouma, Martha Hofman, Jelle Brouwer, Eduard Drenth, Jan Wijffels and Hans Van de Velde . . . . . 4789

*A Dataset of Offensive German Language Tweets Annotated for Speech Acts*
Melina Plakidis and Georg Rehm . . . . . 4799

*Tracing Syntactic Change in the Scientific Genre: Two Universal Dependency-parsed Diachronic Corpora of Scientific English and German*
Marie-Pauline Krielke, Luigi Talamo, Mahmoud Fawzi and Jörg Knappen . . . . . 4808

*The Tembusu Treebank: An English Learner Treebank*
Luís Morgado da Costa, Francis Bond and Roger V. P. Winder . . . . . 4817

*The Norwegian Dialect Corpus Treebank*
Andre Kåsen, Kristin Hagen, Anders Nøklestad, Joel Priestly, Per Erik Solberg and Dag Trygve Truslew Haug . . . . . 4827

*RRGparbank: A Parallel Role and Reference Grammar Treebank*
Tatiana Bladier, Kilian Evang, Valeria Generalova, Zahra Ghane, Laura Kallmeyer, Robin Möllemann, Natalia Moors, Rainer Osswald and Simon Petitjean . . . . . 4833

*Unifying Morphology Resources with OntoLex-Morph. A Case Study in German*
Christian Chiarcos, Christian Fäth and Maxim Ionov . . . . . 4842

*Building Dataset for Grounding of Formulae — Annotating Coreference Relations Among Math Identifiers*
Takuto Asakura, Yusuke Miyao and Akiko Aizawa . . . . . 4851

*CorefUD 1.0: Coreference Meets Universal Dependencies*
Anna Nedoluzhko, Michal Novák, Martin Popel, Zdeněk Žabokrtský, Amir Zeldes and Daniel Zeman . . . . . 4859

*The Universal Anaphora Scorer*
Juntao Yu, Sopan Khosla, Nafise Sadat Moosavi, Silviu Paun, Sameer Pradhan and Massimo Poesio 4873

*Towards Evaluation of Cross-document Coreference Resolution Models Using Datasets with Diverse Annotation Schemes*
Anastasia Zhukova, Felix Hamborg and Bela Gipp . . . . . 4884

*Explainable Tsetlin Machine Framework for Fake News Detection with Credibility Score Assessment*
Bimal Bhattarai, Ole-Christoffer Granmo and Lei Jiao . . . . . 4894

*Enhancing Deep Learning with Embedded Features for Arabic Named Entity Recognition*
Ali L. Hatab, Caroline Sabty and Slim Abdennadher . . . . . 4904

*SCAI-QReCC Shared Task on Conversational Question Answering*
Svitlana Vakulenko, Johannes Kiesel and Maik Fröbe . . . . . 4913

*Semantic Relations between Text Segments for Semantic Storytelling: Annotation Tool - Dataset - Evaluation*
Michael Raring, Malte Ostendorff and Georg Rehm . . . . . 4923

*Evaluating Pre-training Objectives for Low-Resource Translation into Morphologically Rich Languages*
Prajit Dhar, Arianna Bisazza and Gertjan van Noord . . . . . 4933

*Aligning Images and Text with Semantic Role Labels for Fine-Grained Cross-Modal Understanding*
Abhidip Bhattacharyya, Cecilia Mauceri, Martha Palmer and Christoffer Heckman . . . . . 4944

*Rosetta-LSF: an Aligned Corpus of French Sign Language and French for Text-to-Sign Translation*
Elise Bertin-Lemée, Annelies Braffort, Camille Challant, Claire Danet, Boris Dauriac, Michael Filhol, Emmanuella Martinod and Jérémie Segouat . . . . . 4955

*MLQE-PE: A Multilingual Quality Estimation and Post-Editing Dataset*
Marina Fomicheva, Shuo Sun, Erick Fonseca, Chrysoula Zerva, Frédéric Blain, Vishrav Chaudhary, Francisco Guzmán, Nina Lopatina, Lucia Specia and André F. T. Martins . . . . . 4963

*OpenKorPOS: Democratizing Korean Tokenization with Voting-Based Open Corpus Annotation*
Sangwhan Moon, Won Ik Cho, Hye Joo Han, Naoaki Okazaki and Nam Soo Kim . . . . . 4975

*Enriching Grammatical Error Correction Resources for Modern Greek*
Katerina Korre and John Pavlopoulos . . . . . 4984

*A Hmong Corpus with Elaborate Expression Annotations*
David R. Mortensen, Xinyu Zhang, Chenxuan Cui and Katherine Zhang . . . . . 4992

*ELAL: An Emotion Lexicon for the Analysis of Alsatian Theatre Plays*
Delphine Bernhard and Pablo Ruiz Fabo . . . . . 5001

*Universal Dependencies for Western Sierra Puebla Nahuatl*
Robert Pugh, Marivel Huerta Mendez, Mitsuya Sasaki and Francis Tyers . . . . . 5011

*The Construction and Evaluation of the LEAFTOP Dataset of Automatically Extracted Nouns in 1480 Languages*
Gregory Baker and Diego Molla . . . . . 5021

*Huqariq: A Multilingual Speech Corpus of Native Languages of Peru forSpeech Recognition*
Rodolfo Zevallos, Luis Camacho and Nelsi Melgarejo . . . . . 5029

*Writing System and Speaker Metadata for 2,800+ Language Varieties*
Daan van Esch, Tamar Lucassen, Sebastian Ruder, Isaac Caswell and Clara Rivera . . . . . 5035

*The PALMA Corpora of African Varieties of Portuguese*
Tjerk Hagemeijer, Amália Mendes, Rita Gonçalves, Catarina Cornejo, Raquel Madureira and Michel Généreux . . . . . 5047

*A Learning-Based Dependency to Constituency Conversion Algorithm for the Turkish Language*
Büşra Marşan, Oğuz K. Yıldız, Aslı Kuzgun, Neslihan Cesur, Arife B. Yenice, Ezgi Sanıyar, Oğuzhan Kuyrukçu, Bilge N. Arıcan and Olcay Taner Yıldız . . . . . 5054

*Standard German Subtitling of Swiss German TV content: the PASSAGE Project*
Jonathan David Mutal, Pierrette Bouillon, Johanna Gerlach and Veronika Haberkorn . . . . . 5063

*A Survey of Multilingual Models for Automatic Speech Recognition*
Hemant Yadav and Sunayana Sitaram . . . . . 5071

*LuxemBERT: Simple and Practical Data Augmentation in Language Model Pre-Training for Luxembourgish*
Cedric Lothritz, Bertrand Lebichot, Kevin Allix, Lisa Veiber, TEGAWENDE BISSYANDE, Jacques Klein, Andrey Boytsov, Clément Lefebvre and Anne Goujon . . . . . 5080

*PerPaDa: A Persian Paraphrase Dataset based on Implicit Crowdsourcing Data Collection*
Salar Mohtaj, Fatemeh Tavakkoli and Habibollah Asghari . . . . . 5090

*Introducing the Welsh Text Summarisation Dataset and Baseline Systems*
Ignatius Ezeani, Mahmoud El-Haj, Jonathan Morris and Dawn Knight . . . . . 5097

*A Systematic Approach to Derive a Refined Speech Corpus for Sinhala*
Disura Warusawithana, Nilmani Kulaweera, Lakshan Weerasinghe and Buddhika Karunarathne 5107

*IgboBERT Models: Building and Training Transformer Models for the Igbo Language*
Chiamaka Chukwuneke, Ignatius Ezeani, Paul Rayson and Mahmoud El-Haj . . . . . 5114

*Latvian National Corpora Collection – Korpuss.lv*
Baiba Saulite, Roberts Darģis, Normunds Gruzitis, Ilze Auzina, Kristīne Levāne-Petrova, Lauma Pretkalniņa, Laura Rituma, Peteris Paikens, Arturs Znotins, Laine Strankale, Kristīne Pokratniece, Ilmārs Poikāns, Guntis Barzdins, Inguna Skadiņa, Anda Baklāne, Valdis Saulespurēns and Jānis Ziediņš . 5123

*Investigating the Relationship Between Romanian Financial News and Closing Prices from the Bucharest Stock Exchange*
Ioan-Bogdan Iordache, Ana Sabina Uban, Catalin Stoean and Liviu P. Dinu . . . . . 5130

*A Free/Open-Source Morphological Analyser and Generator for Sakha*
Sardana Ivanova, Jonathan Washington and Francis Tyers . . . . . 5137

*An Expanded Finite-State Transducer for Tsuut'ina Verbs*
Joshua Holden, Christopher Cox and Antti Arppe . . . . . 5143

*BD-SHS: A Benchmark Dataset for Learning to Detect Online Bangla Hate Speech in Different Social Contexts*
Nauros Romim, Mosahed Ahmed, Md Saiful Islam, Arnab Sen Sharma, Hriteshwar Talukder and Mohammad Ruhul Amin . . . . . 5153

*Introducing RezoJDM16k: a French KnowledgeGraph DataSet for Link Prediction*
Mehdi Mirzapour, Waleed Ragheb, Mohammad Javad Saeedizade, Kevin Cousot, Helene Jacquenet, Lawrence Carbon and Mathieu Lafourcade . . . . . 5163

*The Badalona Corpus - An Audio, Video and Neuro-Physiological Conversational Dataset*
Philippe Blache, Salomé Antoine, Dorina De Jong, Lena-Marie Huttner, Emilia Kerr, Thierry Legou, Eliot Maës and Clément François . . . . . 5170

*Reading Time and Vocabulary Rating in the Japanese Language: Large-Scale Japanese Reading Time Data Collection Using Crowdsourcing*
Masayuki Asahara . . . . . 5178

*Thematic Fit Bits: Annotation Quality and Quantity Interplay for Event Participant Representation*
Yuval Marton and Asad Sayeed . . . . . 5188

*ChiSense-12: An English Sense-Annotated Child-Directed Speech Corpus*
Francesco Cabiddu, Lewis Bott, Gary Jones and Chiara Gambi . . . . . 5198

*Making People Laugh like a Pro: Analysing Humor Through Stand-Up Comedy*
Beatrice Turano and Carlo Strapparava . . . . . 5206

*Testing Focus and Non-at-issue Frameworks with a Question-under-Discussion-Annotated Corpus*
Christoph Hesse, Maurice Langner, Ralf Klabunde and Anton Benz . . . . . 5212

*Development of a Multilingual CCG Treebank via Universal Dependencies Conversion*
Tu-Anh Tran and Yusuke Miyao . . . . . 5220

*The Automatic Extraction of Linguistic Biomarkers as a Viable Solution for the Early Diagnosis of Mental Disorders*
Gloria Gagliardi and Fabio Tamburini . . . . . 5234

*Singlish Where Got Rules One? Constructing a Computational Grammar for Singlish*
Siew Yeng Chow and Francis Bond . . . . . 5243

*COSMOS: Experimental and Comparative Studies of Concept Representations in Schoolchildren*
Jeanne Villaneau and Farida SAID . . . . . 5251

*Features of Perceived Metaphoricity on the Discourse Level: Abstractness and Emotionality*
Prisca Piccirilli and Sabine Schulte im Walde . . . . . 5261

*Hollywood Identity Bias Dataset: A Context Oriented Bias Analysis of Movie Dialogues*
Sandhya Singh, Prapti Roy, Nihar Sahoo, Niteesh Mallela, Himanshu Gupta, Pushpak Bhattacharyya, Milind Savagaonkar, Nidhi Sultan, Roshni Ramnani, Anutosh Maitra and Shubhashis Sengupta . . . 5274

*VoxCommunis: A Corpus for Cross-linguistic Phonetic Analysis*
Emily Ahn and Eleanor Chodroff . . . . . 5286

*Tracking Textual Similarities in Neo-Latin Drama Networks*
Andrea Peverelli, Marieke van Erp and Jan Bloemendal . . . . . 5295

*Named Entity Recognition in Estonian 19th Century Parish Court Records*
Siim Orasmaa, Kadri Muischnek, Kristjan Poska and Anna Edela . . . . . 5304

*Investigating Independence vs. Control: Agenda-Setting in Russian News Coverage on Social Media*
Annerose Eichel, Gabriella Lapesa and Sabine Schulte im Walde . . . . . 5314

*SLäNDa version 2.0: Improved and Extended Annotation of Narrative and Dialogue in Swedish Literature*
Sara Stymne and Carin Östman . . . . . 5324

*AGILe: The First Lemmatizer for Ancient Greek Inscriptions*
Evelien de Graaf, Silvia Stopponi, Jasper K. Bos, Saskia Peels-Matthey and Malvina Nissim . 5334

*»textklang« – Towards a Multi-Modal Exploration Platform for German Poetry*
Nadja Schauffler, Toni Bernhart, Andre Blessing, Gunilla Eschenbach, Markus Gärtner, Kerstin Jung, Anna Kinder, Julia Koch, Sandra Richter, Gabriel Viehhauser, Ngoc Thang Vu, Lorenz Wesemann and Jonas Kuhn . . . . . 5345

*Predicting the Proficiency Level of Nonnative Hebrew Authors*
Isabelle Nguyen and Shuly Wintner . . . . . 5356

*Trends, Limitations and Open Challenges in Automatic Readability Assessment Research*
Sowmya Vajjala . . . . . 5366

*HateCheckHIn: Evaluating Hindi Hate Speech Detection Models*
Mithun Das, Punyajoy Saha, Binny Mathew and Animesh Mukherjee . . . . . 5378

*Surfer100: Generating Surveys From Web Resources, Wikipedia-style*
Irene Li, Alex Fabbri, Rina Kawamura, Yixin Liu, Xiangru Tang, Jaesung tae, Chang Shen, Sally Ma, Tomoe Mizutani and Dragomir Radev . . . . . 5388

*MS-LaTTE: A Dataset of Where and When To-do Tasks are Completed*
Sujay Kumar Jauhar, Nirupama Chandrasekaran, Michael Gamon and Ryen White . . . . . 5393

*KazakhTTS2: Extending the Open-Source Kazakh TTS Corpus With More Data, Speakers, and Topics*
Saida Mussakhojayeva, Yerbolat Khassanov and Huseyin Atakan Varol . . . . . 5404

*A Graph-Based Method for Unsupervised Knowledge Discovery from Financial Texts*
Joel Oksanen, Abhilash Majumder, Kumar Saunack, Francesca Toni and Arun Dhondiyal . . . 5412

*Leveraging Mental Health Forums for User-level Depression Detection on Social Media*
Sravani Boinepelli, Tathagata Raha, Harika Abburi, Pulkit Parikh, Niyati Chhaya and Vasudeva Varma . . . . . 5418

*Classifying Implant-Bearing Patients via their Medical Histories: a Pre-Study on Swedish EMRs with Semi-Supervised GanBERT*
Benjamin Danielsson, Marina Santini, Peter Lundberg, Yosef Al-Abasse, Arne Jonsson, Emma Eneling and Magnus Stridsman . . . . . 5428

*Standardisation of Dialect Comments in Social Networks in View of Sentiment Analysis : Case of Tunisian Dialect*
Saméh Kchaou, rahma boujelbane, Emna Fsih and Lamia Hadrich-Belguith . . . . . 5436

*EnsyNet: A Dataset for Encouragement and Sympathy Detection*
Tiberiu Sosea and Cornelia Caragea . . . . . 5444

*Preliminary Results on the Evaluation of Computational Tools for the Analysis of Quechua and Aymara*
Marcelo Yuji Himoro and Antonio Pareja-Lora . . . . . 5450

*A Tale of Two Regulatory Regimes: Creation and Analysis of a Bilingual Privacy Policy Corpus*
Siddhant Arora, Henry Hosseini, Christine Utz, Vinayshekhar Bannihatti Kumar, Tristan Dhellemmes, Abhilasha Ravichander, Peter Story, Jasmine Mangat, Rex Chen, Martin Degeling, Thomas Norton, Thomas Hupperich, Shomir Wilson and Norman Sadeh . . . . . 5460

*MeSHup: Corpus for Full Text Biomedical Document Indexing*
Xindi Wang, Robert E. Mercer and Frank Rudzicz . . . . . 5473

*Hierarchical Annotation for Building A Suite of Clinical Natural Language Processing Tasks: Progress Note Understanding*
Yanjun Gao, Dmitriy Dligach, Timothy Miller, Samuel Tesch, Ryan Laffin, Matthew M. Churpek and Majid Afshar . . . . . 5484

*KC4MT: A High-Quality Corpus for Multilingual Machine Translation*
Vinh Van Nguyen, Ha Nguyen, Huong Thanh Le, Thai Phuong Nguyen, Tan Van Bui, Luan Nghia Pham, Anh Tuan Phan, Cong Hoang-Minh Nguyen, Viet Hong Tran and Anh Huu Tran . . . . . 5494

*Developing A Multilabel Corpus for the Quality Assessment of Online Political Talk*
Kokil Jaidka . . . . . 5503

*BILinMID: A Spanish-English Corpus of the US Midwest*
Irati Hurtado . . . . . 5511

*One Document, Many Revisions: A Dataset for Classification and Description of Edit Intents*
Dheeraj Rajagopal, Xuchao Zhang, Michael Gamon, Sujay Kumar Jauhar, Diyi Yang and Eduard Hovy . . . . . 5517

*CTAP for Chinese:A Linguistic Complexity Feature Automatic Calculation Platform*
Yue Cui, Junhui Zhu, Liner Yang, Xuezhi Fang, Xiaobin Chen, Yujie Wang and Erhong Yang 5525

*A Corpus for Suggestion Mining of German Peer Feedback*
Dominik Pfütze, Eva Ritz, Julius Janda and Roman Rietsche . . . . . 5539

*CLGC: A Corpus for Chinese Literary Grace Evaluation*
Yi Li, Dong Yu and pengyuan liu . . . . . 5548

*Anonymising the SAGT Speech Corpus and Treebank*
Özlem Çetinoğlu and Antje Schweitzer . . . . . 5557

*Construction of a Quality Estimation Dataset for Automatic Evaluation of Japanese Grammatical Error Correction*
Daisuke Suzuki, Yujin Takahashi, Ikumi Yamashita, Taichi Aida, Tosho Hirasawa, Michitaka Nakatsuji, Masato Mita and Mamoru Komachi . . . . . 5565

*Enhanced Distant Supervision with State-Change Information for Relation Extraction*
Jui Shah, Dongxu Zhang, Sam Brody and Andrew McCallum . . . . . 5573

*The Hebrew Essay Corpus*
Chen Gafni, Anat Prior and Shuly Wintner . . . . . 5580

*Design and Evaluation of the Corpus of Everyday Japanese Conversation*
Hanae Koiso, Haruka Amatani, Yasuharu Den, Yuriko Iseki, Yuichi Ishimoto, Wakako Kashino, Yoshiko Kawabata, Ken'ya Nishikawa, Yayoi Tanaka, Yasuyuki Usuda and Yuka Watanabe . . . . . 5587

*Developing Language Resources and NLP Tools for the North Korean Language*
Arda Akdemir, Yeojoo Jeon and Tetsuo Shibuya . . . . . 5595

*Developing a Dataset of Overridden Information in Wikipedia*
Masatoshi Tsuchiya and Yasutaka Yokoi . . . . . 5601

*BRATECA (Brazilian Tertiary Care Dataset): a Clinical Information Dataset for the Portuguese Language*
Bernardo Consoli, Henrique D. P. dos Santos, Ana Helena D. P. S. Ulbrich, Renata Vieira and Rafael H. Bordini . . . . . 5609

*Universal Grammatical Dependencies for Portuguese with CINTIL Data, LX Processing and CLARIN support*
António Branco, João Ricardo Silva, Luís Gomes and João António Rodrigues . . . . . 5617

*CWID-hi: A Dataset for Complex Word Identification in Hindi Text*
Gayatri Venugopal, Dhanya Pramod and Ravi Shekhar . . . . . 5627

*Automatic Classification of Russian Learner Errors*
Alla Rozovskaya . . . . . 5637

*Annotation of metaphorical expressions in the Basic Corpus of Polish Metaphors*
Elżbieta Hajnicz . . . . . 5648

*ChiMST: A Chinese Medical Corpus for Word Segmentation and Medical Term Recognition*
Yuanhe Tian, Han Qin, Fei Xia and Yan Song . . . . . 5654

*Building a Synthetic Biomedical Research Article Citation Linkage Corpus*
Sudipta Singha Roy and Robert E. Mercer . . . . . 5665

*Dataset Construction for Scientific-Document Writing Support by Extracting Related Work Section and Citations from PDF Papers*
Keita Kobayashi, Kohei Koyama, Hiromi Narimatsu and Yasuhiro Minami . . . . . 5673

*RuPAWS: A Russian Adversarial Dataset for Paraphrase Identification*
Nikita Martynov, Irina Krotova, Varvara Logacheva, Alexander Panchenko, Olga Kozlova and Nikita Semenov . . . . . 5683

*Atril: an XML Visualization System for Corpus Texts*
Andressa Rodrigues Gomide, Conceição Carapinha and Cornelia Plag . . . . . 5692

*MASALA: Modelling and Analysing the Semantics of Adpositions in Linguistic Annotation of Hindi*
Aryaman Arora, Nitin Venkateswaran and Nathan Schneider . . . . . 5696

*Universal Dependencies for Punjabi*
Aryaman Arora . . . . . 5705

*TeSum: Human-Generated Abstractive Summarization Corpus for Telugu*
Ashok Urlana, Nirmal Surange, Pavan Baswani, Priyanka Ravva and Manish Shrivastava . . . . 5712

*A Corpus of Simulated Counselling Sessions with Dialog Act Annotation*
John Lee, Haley Fong, Lai Shuen Judy Wong, Chun Chung Mak, Chi Hin Yip and Ching Wah Larry Ng . . . . . 5723

*Interactive Evaluation of Dialog Track at DSTC9*
Shikib Mehri, Yulan Feng, Carla Gordon, Seyed Hossein Alavi, David Traum and Maxine Eskenazi 5731

*HADREB: Human Appraisals and (English) Descriptions of Robot Emotional Behaviors*
Josue Torres-Fonsesca and Casey Kennington . . . . . 5739

*Dialogue Collection for Recording the Process of Building Common Ground in a Collaborative Task*
Koh Mitsuda, Ryuichiro Higashinaka, Yuhei Oga and Sen Yoshida . . . . . 5749

*Collection and Analysis of Travel Agency Task Dialogues with Age-Diverse Speakers*
Michimasa Inaba, Yuya Chiba, Ryuichiro Higashinaka, Kazunori Komatani, Yusuke Miyao and Takayuki Nagai . . . . . 5759

*Strategy-level Entrainment of Dialogue System Users in a Creative Visual Reference Resolution Task*
Deepthi Karkada, Ramesh Manuvinakurike, Maike Paetzel-Prüsmann and Kallirroi Georgila . 5768

*MMChat: Multi-Modal Chat Dataset on Social Media*
Yinhe Zheng, Guanyi Chen, Xin Liu and Jian Sun . . . . . 5778

*E-ConvRec: A Large-Scale Conversational Recommendation Dataset for E-Commerce Customer Service*
meihuizi jia, Ruixue Liu, Peiying Wang, Yang Song, Zexi Xi, Haobin Li, Xin Shen, Meng Chen, Jinhui Pang and Xiaodong He . . . . . 5787

*SHONGLAP: A Large Bengali Open-Domain Dialogue Corpus*
Syed Mostofa Monsur, Sakib Chowdhury, Md Shahrar Fatemi and Shafayat Ahmed . . . . . 5797

*A Comparison of Praising Skills in Face-to-Face and Remote Dialogues*
Toshiki Onishi, Asahi Ogushi, Yohei Tahara, Ryo Ishii, Atsushi Fukayama, Takao Nakamura and Akihiro Miyata . . . . . 5805

*Comparing Approaches to Language Understanding for Human-Robot Dialogue: An Error Taxonomy and Analysis*
Ada Tur and David Traum . . . . . 5813

*SPORTSINTERVIEW: A Large-Scale Sports Interview Benchmark for Entity-centric Dialogues*
Hanfei Sun, Ziyuan Cao and Diyi Yang . . . . . 5821

*EmoInHindi: A Multi-label Emotion and Intensity Annotated Dataset in Hindi for Emotion Recognition in Dialogues*
Gopendra Vikram Singh, Priyanshu Priya, Mauajama Firdaus, Asif Ekbal and Pushpak Bhattacharyya . . . . . 5829

*The Project Dialogism Novel Corpus: A Dataset for Quotation Attribution in Literary Texts*
Krishnapriya Vishnubhotla, Adam Hammond and Graeme Hirst . . . . . 5838

*Who's in, who's out? Predicting the Inclusiveness or Exclusiveness of Personal Pronouns in Parliamentary Debates*
Ines Rehbein and Josef Ruppenhofer . . . . . 5849

*A Language Modelling Approach to Quality Assessment of OCR'ed Historical Text*
Callum Booth, Robert Shoemaker and Robert Gaizauskas . . . . . 5859

*Identifying Copied Fragments in a 18th Century Dutch Chronicle*
Roser Morante, Eleanor L. T. Smith, Lianne Wilhelmus, Alie Lassche and Erika Kuijpers . . . . 5865

*A Study of Distant Viewing of ukiyo-e prints*
Konstantina Liagkou, John Pavlopoulos and Ewa Machotka . . . . . 5879

*CCTAA: A Reproducible Corpus for Chinese Authorship Attribution Research*
Haining Wang and Allen Riddell . . . . . 5889

*An automatic model and Gold Standard for translation alignment of Ancient Greek*
Tariq Yousef, Chiara Palladino, Farnoosh Shamsian, Anise d'Orange Ferreira and Michel Ferreira dos Reis . . . . . 5894

*Rhetorical Structure Approach for Online Deception Detection: A Survey*
Francielle Vargas, Jonas D'Alessandro, Zohar Rabinovich, Fabrício Benevenuto and Thiago Pardo 5906

*TYPIC: A Corpus of Template-Based Diagnostic Comments on Argumentation*
Shoichi Naito, Shintaro Sawada, Chihiro Nakagawa, Naoya Inoue, Kenshi Yamaguchi, Iori Shimizu, Farjana Sultana Mim, Keshav Singh and Kentaro Inui . . . . . 5916

*Towards Speaker Verification for Crowdsourced Speech Collections*
John Mendonca, Rui Correia, Mariana Lourenço, João Freitas and Isabel Trancoso . . . . . 5929

*Align-smatch: A Novel Evaluation Method for Chinese Abstract Meaning Representation Parsing based on Alignment of Concept and Relation*
Liming Xiao, Bin Li, Zhixing Xu, Kairui Huo, Minxuan Feng, Junsheng Zhou and Weiguang Qu 5938

*Dynamic Human Evaluation for Relative Model Comparisons*
Thórhildur Thorleiksdóttir, Cedric Renggli, Nora Hollenstein and Ce Zhang . . . . . 5946

*Please, Don't Forget the Difference and the Confidence Interval when Seeking for the State-of-the-Art Status*
Yves Bestgen . . . . . 5956

*PCR4ALL: A Comprehensive Evaluation Benchmark for Pronoun Coreference Resolution in English*
Xinran Zhao, Hongming Zhang and Yangqiu Song . . . . . 5963

*Estimating Confidence of Predictions of Individual Classifiers and TheirEnsembles for the Genre Classification Task*
Mikhail Lepekhin and Serge Sharoff . . . . . 5974

*What do we really know about State of the Art NER?*
Sowmya Vajjala and Ramya Balasubramaniam . . . . . 5983

*ProQE: Proficiency-wise Quality Estimation dataset for Grammatical Error Correction*
Yujin Takahashi, Masahiro Kaneko, Masato Mita and Mamoru Komachi . . . . . 5994

*Evaluation of Off-the-shelf Speech Recognizers on Different Accents in a Dialogue Domain*
Divya Tadimeti, Kallirroi Georgila and David Traum . . . . . 6001

*Sentence Pair Embeddings Based Evaluation Metric for Abstractive and Extractive Summarization*
Ramya Akula and Ivan Garibay . . . . . 6009

*On "Human Parity" and "Super Human Performance" in Machine Translation Evaluation*
Thierry Poibeau . . . . . 6018

*Evaluation Benchmarks for Spanish Sentence Representations*
Vladimir Araujo, Andrés Carvallo, Souvik Kundu, José Cañete, Marcelo Mendoza, Robert E. Mercer, Felipe Bravo-Marquez, Marie-Francine Moens and Alvaro Soto . . . . . 6024

*UMUTextStats: A linguistic feature extraction tool for Spanish*
José Antonio García-Díaz, Pedro José Vivancos-Vicente, Ángela Almela and Rafael Valencia-García . . . . . 6035

*Problem-solving Recognition in Scientific Text*
Kevin Heffernan and Simone Teufel . . . . . 6045

*HRCA+: Advanced Multiple-choice Machine Reading Comprehension Method*
YUXIANG ZHANG and Hayato Yamana . . . . . 6059

*HyperBox: A Supervised Approach for Hypernym Discovery using Box Embeddings*
Maulik Parmar and Apurva Narayan . . . . . 6069

*Extracting Space Situational Awareness Events from News Text*
Zhengnan Xie, Alice Saebom Kwak, Enfa George, Laura W. Dozal, Hoang Van, Moriba Jah, Roberto Furfaro and Peter Jansen . . . . . 6077

*PerCQA: Persian Community Question Answering Dataset*
Naghme Jamali, Yadollah Yaghoobzadeh and Heshaam Faili . . . . . 6083

*GrASP: A Library for Extracting and Exploring Human-Interpretable Textual Patterns*
Piyawat Lertvittayakumjorn, Leshem Choshen, Eyal Shnarch and Francesca Toni . . . . . 6093

*Recurrent Neural Networks with Mixed Hierarchical Structures and EM Algorithm for Natural Language Processing*
zhaoxin luo and Michael Zhu . . . . . 6104

*Korean-Specific Dataset for Table Question Answering*
Changwook Jun, Jooyoung Choi, Myoseop Sim, Hyun Kim, Hansol Jang and Kyungkoo Min 6114

*GerCCT: An Annotated Corpus for Mining Arguments in German Tweets on Climate Change*
Robin Schaefer and Manfred Stede . . . . . 6121

*Budget Argument Mining Dataset Using Japanese Minutes from the National Diet and Local Assemblies*
Yasutomo Kimura, Hokuto Ototake and Minoru Sasaki . . . . . 6131

*Context-based Virtual Adversarial Training for Text Classification with Noisy Labels*
Do-Myoung Lee, Yeachan Kim and Chang gyun Seo . . . . . 6139

*FinMath: Injecting a Tree-structured Solver for Question Answering over Financial Reports*
Chenying Li, Wenbo Ye and Yilun Zhao . . . . . 6147

*HeadlineCause: A Dataset of News Headlines for Detecting Causalities*
Ilya Gusev and Alexey Tikhonov . . . . . 6153

*Incorporating Zoning Information into Argument Mining from Biomedical Literature*
Boyang Liu, Viktor Schlegel, Riza Batista-Navarro and Sophia Ananiadou . . . . . 6162

*MAKED: Multi-lingual Automatic Keyword Extraction Dataset*
Yash Verma, Anubhav Jangra, Sriparna Saha, Adam Jatowt and Dwaipayan Roy . . . . . 6170

*From Examples to Rules: Neural Guided Rule Synthesis for Information Extraction*
Robert Vacareanu, Marco A. Valenzuela-Escárcega, George Caique Gouveia Barbosa, Rebecca Sharp, Gustave Hahn-Powell and Mihai Surdeanu . . . . . 6180

*Enhancing Relation Extraction via Adversarial Multi-task Learning*
Han Qin, Yuanhe Tian and Yan Song . . . . . 6190

*Query Obfuscation by Semantic Decomposition*
Danushka Bollegala, Tomoya Machide and Ken-ichi Kawarabayashi . . . . . 6200

*TWEET-FID: An Annotated Dataset for Multiple Foodborne Illness Detection Tasks*
Ruofan Hu, Dongyu Zhang, Dandan Tao, Thomas Hartvigsen, Hao Feng and Elke Rundensteiner 6212

*Named Entity Recognition to Detect Criminal Texts on the Web*
Paweł Skórzewski, Mikołaj Pieniowski and Grazyna Demenko . . . . . 6223

*Task-Driven and Experience-Based Question Answering Corpus for In-Home Robot Application in the House3D Virtual Environment*
zhuoqun Xu, Liubo Ouyang and Yang Liu . . . . . 6232

*ELRC Action: Covering Confidentiality, Correctness and Cross-linguality*
Tom Vanallemeersch, Arne Defauw, Sara Szoc, Alina Kramchaninova, Joachim Van den Bogaert and Andrea Lösch . . . . . 6240

*RadQA: A Question Answering Dataset to Improve Comprehension of Radiology Reports*
Sarvesh Soni, Meghana Gudala, Atieh Pajouhi and Kirk Roberts . . . . . 6250

*Knowledge Graph - Deep Learning: A Case Study in Question Answering in Aviation Safety Domain*
Ankush Agarwal, Raj Gite, Shreya Laddha, Pushpak Bhattacharyya, Satyanarayan Kar, Asif Ekbal, Prabhjit Thind, Rajesh Zele and Ravi Shankar . . . . . 6260

*A Bayesian Topic Model for Human-Evaluated Interpretability*
Justin Wood, Corey Arnold and Wei Wang . . . . . 6271

*A Large Interlinked Knowledge Graph of the Italian Cultural Heritage*
Stefano Faralli, Andrea Lenzi and Paola Velardi . . . . . 6280

*Training on Lexical Resources*
Kenneth Church, Xingyu Cai and Yuchen Bian . . . . . 6290

*Challenging the Assumption of Structure-based embeddings in Few- and Zero-shot Knowledge Graph Completion*
Filip Cornell, Chenda zhang, Jussi Karlgren and Sarunas Girdzijauskas . . . . . 6300

*Open Terminology Management and Sharing Toolkit for Federation of Terminology Databases*
Andis Lagzdiņš, Uldis Siliņš, Toms Bergmanis, Mārcis Pinnis, Artūrs Vasiļevskis and Andrejs Vasiljevs . . . . . 6310

*RELATE: Generating a linguistically inspired Knowledge Graph for fine-grained emotion classification*
Annika Marie Schoene, Nina Dethlefs and Sophia Ananiadou . . . . . 6317

*Language technology practitioners as language managers: arbitrating data bias and predictive bias in ASR*
Nina Markl and Stephen Joseph McNulty . . . . . 6328

*Masader: Metadata Sourcing for Arabic Text and Speech Data Resources*
Zaid Alyafeai, Maraim Masoud, Mustafa Ghaleb and Maged S. Al-shaibani . . . . . 6340

*Linghub2: Language Resource Discovery Tool for Language Technologies*
Cécile Robin, Gautham Vadakkekara Suresh, Víctor Rodriguez-Doncel, John P. McCrae and Paul Buitelaar . . . . . 6352

*CxLM: A Construction and Context-aware Language Model*
Yu-Hsiang Tseng, Cing-Fang Shih, Pin-Er Chen, Hsin-Yu Chou, Mao-Chang Ku and Shu-Kai HSIEH . . . . . 6361

*The Lexometer: A Shiny Application for Exploratory Analysis and Visualization of Corpus Data*
Oufan Hai, Matthew Sundberg, Katherine Trice, Rebecca Friedman and Scott Grimm . . . . . 6370

*TallVocabL2Fi: A Tall Dataset of 15 Finnish L2 Learners' Vocabulary*
Frankie Robertson, Li-Hsin Chang and Sini Söyrinki . . . . . 6377

*CAMS: An Annotated Corpus for Causal Analysis of Mental Health Issues in Social Media Posts*
Muskan Garg, Chandni Saxena, Sriparna Saha, Veena Krishnan, Ruchi Joshi and Vijay Mago 6387

*How Does the Experimental Setting Affect the Conclusions of Neural Encoding Models?*
Xiaohan Zhang, Shaonan Wang and Chengqing Zong . . . . . 6397

*SPADE: A Big Five-Mturk Dataset of Argumentative Speech Enriched with Socio-Demographics for Personality Detection*
Elma Kerz, Yu Qiao, Sourabh Zanwar and Daniel Wiechmann . . . . . 6405

*Progress in Multilingual Speech Recognition for Low Resource Languages Kurmanji Kurdish, Cree and Inuktut*
vishwa gupta and Gilles Boulianne . . . . . 6420

*Efficient Entity Candidate Generation for Low-Resource Languages*
Alberto Garcia-Duran, Akhil Arora and Robert West . . . . . 6429

*What a Creole Wants, What a Creole Needs*
Heather Lent, Kelechi Ogueji, Miryam de Lhoneux, Orevaoghene Ahia and Anders Søgaard . 6439

*Extensions to Brahmic script processing within the Nisaba library: new scripts, languages and utilities*
Alexander Gutkin, Cibu Johny, Raiomond Doctor, Lawrence Wolf-Sonkin and Brian Roark . . 6450

*Predicting Embedding Reliability in Low-Resource Settings Using Corpus Similarity Measures*
Jonathan Dunn, Haipeng Li and Damian Sastre . . . . . 6461

*Hausa Visual Genome: A Dataset for Multi-Modal English to Hausa Machine Translation*
Idris Abdulmumin, Satya Ranjan Dash, Musa Abdullahi Dawud, Shantipriya Parida, Shamsuddeen Muhammad, Ibrahim Sa'id Ahmad, Subhadarshi Panda, Ondřej Bojar, Bashir Shehu Galadanci and Bello Shehu Bello . . . . . 6471

*A Survey of Machine Translation Tasks on Nigerian Languages*
Ebelechukwu Nwafor and Anietie Andy . . . . . 6480

*Automatic Speech Recognition Datasets in Cantonese: A Survey and New Dataset*
Tiezheng Yu, Rita Frieske, Peng Xu, Samuel Cahyawijaya, Cheuk Tung YIU, Holy Lovenia, Wenliang Dai, Elham J. Barezi, Qifeng Chen, Xiaojuan Ma, Bertram Shi and Pascale Fung . . . . . 6487

*Survey on Thai NLP Language Resources and Tools*
Ratchakrit Arreerard, Stephen Mander and Scott Piao . . . . . 6495

*LaoPLM: Pre-trained Language Models for Lao*
Nankai Lin, Yingwen Fu, Chuwei Chen, Ziyu Yang and Shengyi JIANG . . . . . 6506

*The Maaloula Aramaic Speech Corpus (MASC): From Printed Material to a Lemmatized and Time-Aligned Corpus*
Ghattas Eid, Esther Seyffarth and Ingo Plag . . . . . 6513

*VIMQA: A Vietnamese Dataset for Advanced Reasoning and Explainable Multi-hop Question Answering*
Khang Le, Hien Nguyen, Tung Le Thanh and Minh Nguyen . . . . . 6521

*Language Identification for Austronesian Languages*
Jonathan Dunn and Wikke Nijhof . . . . . 6530

*A Mapudüngun FST Morphological Analyser and its Web Interface*
Andrés Chandía . . . . . 6540

*Improving Large-scale Language Models and Resources for Filipino*
Jan Christian Blaise Cruz and Charibeth Cheng . . . . . 6548

*Thirumurai: A Large Dataset of Tamil Shaivite Poems and Classification of Tamil Pann*
Shankar Mahadevan, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Prabakaran Chandran, Ruba Priyadharshini, Sangeetha S and Bharathi Raja Chakravarthi . . . . . 6556

*Generating Monolingual Dataset for Low Resource Language Bodo from old books using Google Keep*
Sanjib Narzary, Maharaj Brahma, Mwnthai Narzary, Gwmsrang Muchahary, Pranav Kumar Singh, Apurbalal Senapati, Sukumar Nandi and Bidisha Som . . . . . 6563

*AsNER - Annotated Dataset and Baseline for Assamese Named Entity recognition*
Dhrubajyoti Pathak, Sukumar Nandi and Priyankoo Sarmah . . . . . 6571

*GeezSwitch: Language Identification in Typologically Related Low-resourced East African Languages*
Fitsum Gaim, Wonsuk Yang and Jong C. Park . . . . . 6578

*Handwritten Paleographic Greek Text Recognition: A Century-Based Approach*
Paraskevi Platanou, John Pavlopoulos and Georgios Papaioannou . . . . . 6585

*Quality Control for Crowdsourced Bilingual Dictionary in Low-Resource Languages*
Hiroki Chida, Yohei Murakami and Mondheera Pituxcoosuvarn . . . . . 6590

*An Inflectional Database for Gitksan*
Bruce Oliver, Clarissa Forbes, Changbing Yang, Farhan Samir, Edith Coates, Garrett Nicolai and Miikka Silfverberg . . . . . 6597

*PyCantonese: Cantonese Linguistics and NLP in Python*
Jackson Lee, Litong Chen, Charles Lam, Chaak Ming Lau and Tsz-Him Tsui . . . . . 6607

*Afaan Oromo Hate Speech Detection and Classification on Social Media*
Teshome Mulugeta Ababu and Michael Melese Woldeyohannis . . . . . 6612

*Cross-lingual Linking of Automatically Constructed Frames and FrameNet*
Ryohei Sasano . . . . . 6620

*Aligning the Romanian Reference Treebank and the Valence Lexicon of Romanian Verbs*
Ana-Maria Barbu, Verginica Barbu Mititelu and Cătălin Mititelu . . . . . 6626

*PortiLexicon-UD: a Portuguese Lexical Resource according to Universal Dependencies Model*
Lucelene Lopes, Magali Duran, Paulo Fernandes and Thiago Pardo . . . . . 6635

*Extended Parallel Corpus for Amharic-English Machine Translation*
Andargachew Mekonnen Gezmu, Andreas Nürnberger and Tesfaye Bayu Bati . . . . . 6644

*Low-resource Neural Machine Translation: Benchmarking State-of-the-art Transformer for Wolof<->French*
Cheikh M. Bamba Dione, Alla LO, Elhadji Mamadou Nguer and sileye ba . . . . . 6654

*Criteria for Useful Automatic Romanization in South Asian Languages*
Isin Demirsahin, Cibu Johny, Alexander Gutkin and Brian Roark . . . . . 6662

*BERTology for Machine Translation: What BERT Knows about Linguistic Difficulties for Translation*
Yuqian Dai, Marc de Kamps and Serge Sharoff . . . . . 6674

*CVSS Corpus and Massively Multilingual Speech-to-Speech Translation*
Ye Jia, Michelle Tadmor Ramanovich, Quan Wang and Heiga Zen . . . . . 6691

*JParaCrawl v3.0: A Large-scale English-Japanese Parallel Corpus*
Makoto Morishita, Katsuki Chousa, Jun Suzuki and Masaaki Nagata . . . . . 6704

*Learning How to Translate North Korean through South Korean*
Hwichan Kim, Sangwhan Moon, Naoaki Okazaki and Mamoru Komachi . . . . . 6711

*FGraDA: A Dataset and Benchmark for Fine-Grained Domain Adaptation in Machine Translation*
Wenhao Zhu, Shujian Huang, Tong Pu, Pingxuan Huang, xu zhang, Jian Yu, Wei Chen, Yanfeng Wang and Jiajun CHEN . . . . . 6719

*SansTib, a Sanskrit - Tibetan Parallel Corpus and Bilingual Sentence Embedding Model*
Sebastian Nehrdich . . . . . 6728

*VISA: An Ambiguous Subtitles Dataset for Visual Scene-aware Machine Translation*
Yihang Li, Shuichiro Shimizu, Weiqi Gu, Chenhui Chu and Sadao Kurohashi . . . . . 6735

*A Benchmark Dataset for Multi-Level Complexity-Controllable Machine Translation*
Kazuki Tani, Ryoya Yuasa, Kazuki Takikawa, Akihiro Tamura, Tomoyuki Kajiwara, Takashi Ninomiya and Tsuneo Kato . . . . . 6744

*gaHealth: An English–Irish Bilingual Corpus of Health Data*
Séamus Lankford, Haithem Afli, Órla Ní Loinsigh and Andy Way . . . . . 6753

*Translation Memories as Baselines for Low-Resource Machine Translation*
Rebecca Knowles and Patrick Littell . . . . . 6759

*N24News: A New Dataset for Multimodal News Classification*
Zhen Wang, Xu Shan, Xiangxie Zhang and Jie Yang . . . . . 6768

*MultiSubs: A Large-scale Multimodal and Multilingual Dataset*
Josiah Wang, Josiel Figueiredo and Lucia Specia . . . . . 6776

*CI-AVSR: A Cantonese Audio-Visual Speech Datasetfor In-car Command Recognition*
Wenliang Dai, Samuel Cahyawijaya, Tiezheng Yu, Elham J. Barezi, Peng Xu, Cheuk Tung YIU, Rita Frieske, Holy Lovenia, Genta Winata, Qifeng Chen, Xiaojuan Ma, Bertram Shi and Pascale Fung 6786

*Multimodal Negotiation Corpus with Various Subjective Assessments for Social-Psychological Outcome Prediction from Non-Verbal Cues*
Nobukatsu Hojo, Satoshi Kobashikawa, Saki Mizuno and Ryo Masumura . . . . . 6794

*MMDAG: Multimodal Directed Acyclic Graph Network for Emotion Recognition in Conversation*
Shuo Xu, Yuxiang Jia, Changyong Niu and Hongying Zan . . . . . 6802

*Automatic Gloss-level Data Augmentation for Sign Language Translation*
Jin Yea Jang, Han-Mu Park, Saim Shin, Suna Shin, Byungcheon Yoon and Gahgene Gweon . 6808

*Image Description Dataset for Language Learners*
Kento Tanaka, Taichi Nishimura, Hiroaki Nanjo, Keisuke Shirai, Hirotaka Kameko and Masatake Dantsuji . . . . . 6814

*The Multimodal Annotation Software Tool (MAST)*
Bruno Cardoso and Neil Cohn . . . . . 6822

*A Multimodal German Dataset for Automatic Lip Reading Systems and Transfer Learning*
Gerald Schwiebert, Cornelius Weber, Leyuan Qu, Henrique Siqueira and Stefan Wermter . . . . 6829

*Multimodality for NLP-Centered Applications: Resources, Advances and Frontiers*
Muskan Garg, Seema Wazarkar, Muskaan Singh and Ondřej Bojar . . . . . 6837

*Cross-lingual and Multilingual CLIP*
Fredrik Carlsson, Philipp Eisen, Faton Rekathati and Magnus Sahlgren . . . . . 6848

*BAN-Cap: A Multi-Purpose English-Bangla Image Descriptions Dataset*
Mohammad Faiyaz Khan, S.M. Sadiq-Ur-Rahman Shifath and Md Saiful Islam . . . . . 6855

*SSR7000: A Synchronized Corpus of Ultrasound Tongue Imaging for End-to-End Silent Speech Recognition*
Naoki Kimura, Zixiong Su, Takaaki Saeki and Jun Rekimoto . . . . . 6866

*A Simple Yet Effective Corpus Construction Method for Chinese Sentence Compression*
Yang Zhao, Hiroshi Kanayama, Issei Yoshida, Masayasu Muraoka and Akiko Aizawa . . . . . 6874

*JADE: Corpus for Japanese Definition Modelling*
Han Huang, Tomoyuki Kajiwara and Yuki Arase . . . . . 6884

*Unraveling the Mystery of Artifacts in Machine Generated Text*
Jiashu Pu, Ziyi Huang, Yadong Xi, Guandan Chen, Weijie Chen and Rongsheng Zhang . . . . . 6889

*Logic-Guided Message Generation from Raw Real-Time Sensor Data*
Ernie Chang, Alisa Kovtunova, Stefan Borgwardt, Vera Demberg, Kathryn Chapman and Hui-Syuan Yeh . . . . . 6899

*The Bull and the Bear: Summarizing Stock Market Discussions*
Ayush Kumar, Dhyey Jani, Jay Shah, Devanshu Thakar, Varun Jain and Mayank Singh . . . . . 6909

*Combination of Contextualized and Non-Contextualized Layers for Lexical Substitution in French*
Kévin Espasa, Emmanuel Morin and Olivier Hamon . . . . . 6914

*SuMe: A Dataset Towards Summarizing Biomedical Mechanisms*
Mohaddeseh Bastan, Nishant Shankar, Mihai Surdeanu and Niranjan Balasubramanian . . . . . 6922

*CATAMARAN: A Cross-lingual Long Text Abstractive Summarization Dataset*
zheng chen and Hongyu Lin . . . . . 6932

*Emotion analysis and detection during COVID-19*
Tiberiu Sosea, Chau Pham, Alexander Tekle, Cornelia Caragea and Junyi Jessy Li . . . . . 6938

*Cross-lingual Emotion Detection*
Sabit Hassan, Shaden Shaar and Kareem Darwish . . . . . 6948

*DirectQuote: A Dataset for Direct Quotation Extraction and Attribution in News Articles*
Yuanchi Zhang and Yang Liu . . . . . 6959

*VaccineLies: A Natural Language Resource for Learning to Recognize Misinformation about the COVID-19 and HPV Vaccines*
Maxwell Weinzierl and Sanda Harabagiu . . . . . 6967

*Tackling Irony Detection using Ensemble Classifiers*
Christoph Turban and Udo Kruschwitz . . . . . 6976

*Automatic Construction of an Annotated Corpus with Implicit Aspects*
Aye Aye Mar and Kiyoaki Shirai . . . . . 6985

*A Multimodal Corpus for Emotion Recognition in Sarcasm*
Anupama Ray, Shubham Mishra, Apoorva Nunna and Pushpak Bhattacharyya . . . . . 6992

*Annotation of Valence Unfolding in Spoken Personal Narratives*
Aniruddha Tammewar, Franziska Braun, Gabriel Roccabruna, Sebastian Bayerl, Korbinian Riedhammer and Giuseppe Riccardi . . . . . 7004

*A Large-Scale Japanese Dataset for Aspect-based Sentiment Analysis*
Yuki Nakayama, Koji Murakami, Gautam Kumar, Sudha Bhingardive and Ikuko Hardaway . . 7014

*A Japanese Dataset for Subjective and Objective Sentiment Polarity Classification in Micro Blog Domain*
Haruya Suzuki, Yuto Miyauchi, Kazuki Akiyama, Tomoyuki Kajiwara, Takashi Ninomiya, Noriko Takemura, Yuta Nakashima and Hajime Nagahara . . . . . 7022

*Complementary Learning of Aspect Terms for Aspect-based Sentiment Analysis*
Han Qin, Yuanhe Tian, Fei Xia and Yan Song . . . . . 7029

*Deep One-Class Hate Speech Detection Model*
saugata bose and Dr. Guoxin Su . . . . . 7040

*Opinions in Interactions : New Annotations of the SEMAINE Database*
Valentin Barriere, Slim Essid and Chloé Clavel . . . . . 7049

*Pars-ABSA: a Manually Annotated Aspect-based Sentiment Analysis Benchmark on Farsi Product Reviews*
Taha Shangipour ataei, Kamyar Darvishi, Soroush Javdan, Behrouz Minaei-Bidgoli and Sauleh Eetemadi . . . . . 7056

*HindiMD: A Multi-domain Corpora for Low-resource Sentiment Analysis*
Mamta ., Asif Ekbal, Pushpak Bhattacharyya, Tista Saha, Alka Kumar and Shikha Srivastava 7061

*Sentiment Analysis of Homeric Text: The 1st Book of Iliad*
John Pavlopoulos, Alexandros Xenos and Davide Picca . . . . . 7071

*The Persian Dependency Treebank Made Universal*
Pegah Safari, Mohammad Sadegh Rasooli, Amirsaeid Moloodi and Alireza Nourian . . . . . 7078

*GujMORPH - A Dataset for Creating Gujarati Morphological Analyzer*
Jatayu Baxi and brijesh bhatt . . . . . 7088

*Informal Persian Universal Dependency Treebank*
Roya Kabiri, Simin Karimi and Mihai Surdeanu . . . . . 7096

*Automatic Correction of Syntactic Dependency Annotation Differences*
Andrew Zupon, Andrew Carnie, Michael Hammond and Mihai Surdeanu . . . . . 7106

*Building Large-Scale Japanese Pronunciation-Annotated Corpora for Reading Heteronymous Logograms*
Fumikazu Sato, Naoki Yoshinaga and Masaru Kitsuregawa . . . . . 7113

*StyleKQC: A Style-Variant Paraphrase Corpus for Korean Questions and Commands*
Won Ik Cho, Sangwhan Moon, Jongin Kim, Seokmin Kim and Nam Soo Kim . . . . . 7122

*Syntax-driven Approach for Semantic Role Labeling*
Yuanhe Tian, Han Qin, Fei Xia and Yan Song . . . . . 7129

*HerBERT Based Language Model Detects Quantifiers and Their Semantic Properties in Polish*
Marcin Woliński, Bartłomiej Nitoń, Witold Kieraś and Jakub Szymanik . . . . . 7140

*Lexical Resource Mapping via Translations*
hongchang Bao, Bradley Hauer and Grzegorz Kondrak . . . . . 7147

*Unsupervised Attention-based Sentence-Level Meta-Embeddings from Contextualised Language Models*
Keigo Takahashi and Danushka Bollegala . . . . . 7155

*Identification of Fine-Grained Location Mentions in Crisis Tweets*
Sarthak Khanal, Maria Traskowsky and Doina Caragea . . . . . 7164

*HateBR: A Large Expert Annotated Corpus of Brazilian Instagram Comments for Offensive Language and Hate Speech Detection*
Francielle Vargas, Isabelle Carvalho, Fabiana Rodrigues de Góes, Thiago Pardo and Fabrício Benevenuto . . . . . 7174

*MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare*
Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari and Erik Cambria . . . . . 7184

*Leveraging Hashtag Networks for Multimodal Popularity Prediction of Instagram Posts*
Yu Yun Liao . . . . . 7191

*Annotating the Tweebank Corpus on Named Entity Recognition and Building NLP Models for Social Media Analysis*
Hang Jiang, Yining Hua, Doug Beeferman and Deb Roy . . . . . 7199

*Did that happen? Predicting Social Media Posts that are Indicative of what happened in a scene: A case study of a TV show*
Anietie Andy, Reno Kriz, Sharath Chandra Guntuku, Derry Tanti Wijaya and Chris Callison-Burch 7209

*HashSet - A Dataset For Hashtag Segmentation*
Prashant Kodali, Akshala Bhatnagar, Naman Ahuja, Manish Shrivastava and Ponnurangam Kumaraguru . . . . . 7215

*Using Convolution Neural Network with BERT for Stance Detection in Vietnamese*
Oanh Tran, Anh Cong Phung and Bach Xuan Ngo . . . . . 7220

*Annotation-Scheme Reconstruction for "Fake News" and Japanese Fake News Dataset*
Taichi Murayama, Shohei Hisada, Makoto Uehara, Shoko Wakamiya and Eiji ARAMAKI . . . 7226

*RoBERTuito: a pre-trained language model for social media text in Spanish*
Juan Manuel Pérez, Damián Ariel Furman, Laura Alonso Alemany and Franco M. Luque . . . . 7235

*Construction of Responsive Utterance Corpus for Attentive Listening Response Production*
Koichiro Ito, Masaki Murata, Tomohiro Ohno and Shigeki Matsubara . . . . . 7244

*Speak: A Toolkit Using Amazon Mechanical Turk to Collect and Validate Speech Audio Recordings*
Christopher Song, David Harwath, Tuka Alhanai and James Glass . . . . . 7253

*ASCEND: A Spontaneous Chinese-English Dataset for Code-switching in Multi-turn Conversation*
Holy Lovenia, Samuel Cahyawijaya, Genta Winata, Peng Xu, Yan Xu, Zihan Liu, Rita Frieske, Tiezheng Yu, Wenliang Dai, Elham J. Barezi, Qifeng Chen, Xiaojuan Ma, Bertram Shi and Pascale Fung 7259

*A Romanization System and WebMAUS Aligner for Arabic Varieties*
Jalal Al-Tamimi, Florian Schiel, Ghada Khattab, Navdeep Sokhey, Djegdjiga Amazouz, Abdulrahman Dallak and Hajar Moussa . . . . . 7269

*BembaSpeech: A Speech Recognition Corpus for the Bemba Language*
Claytone Sikasote and Antonios Anastasopoulos . . . . . 7277

*BehanceCC: A ChitChat Detection Dataset For Livestreaming Video Transcripts*
Viet Lai, Amir Pouran Ben Veyseh, Franck Dernoncourt and Thien Huu Nguyen . . . . . 7284

*Adversarial Speech Generation and Natural Speech Recovery for Speech Content Protection*
Sheng Li, Jiyi Li, Qianying Liu and Zhuo Gong . . . . . 7291

*A new European Portuguese corpus for the study of Psychosis through speech analysis*
Maria Forjó, Daniel Neto, Alberto Abad, HSofia Pinto and Joaquim Gago . . . . . 7298

*Investigating Inter- and Intra-speaker Voice Conversion using Audiobooks*
Aghilas SINI, Damien Lolive, Nelly Barbot and Pierre Alain . . . . . 7305

*Multilingual Transfer Learning for Children Automatic Speech Recognition*
Thomas Rolland, Alberto Abad, Catia Cucchiarini and Helmer Strik . . . . . 7314

*BehanceQA: A New Dataset for Identifying Question-Answer Pairs in Video Transcripts*
Amir Pouran Ben Veyseh, Viet Lai, Franck Dernoncourt and Thien Huu Nguyen . . . . . 7321

*Bidirectional Skeleton-Based Isolated Sign Recognition using Graph Convolutional Networks*
Konstantinos M. Dafnis, Evgenia Chroni, Carol Neidle and Dimitri Metaxas . . . . . 7328

*Deep learning-based end-to-end spoken language identification system for domain-mismatched scenario*
Woohyun Kang, Md Jahangir Alam and Abderrahim Fathan . . . . . 7339

*Handwritten Character Generation using Y-Autoencoder for Character Recognition Model Training*
Tomoki Kitagawa, Chee Siang Leow and Hiromitsu Nishizaki . . . . . 7344

*Attention is All you Need for Robust Temporal Reasoning*
Lis Kanashiro Pereira . . . . . 7352

*PoliBERTweet: A Pre-trained Language Model for Analyzing Political Content on Twitter*
Kornraphop Kawintiranon and Lisa Singh . . . . . 7360

*Modeling the Impact of Syntactic Distance and Surprisal on Cross-Slavic Text Comprehension*
Irina Stenger, Philip Georgis, Tania Avgustinova, Bernd Möbius and Dietrich Klakow . . . . . 7368

*BERTifying Sinhala - A Comprehensive Analysis of Pre-trained Language Models for Sinhala Text Classification*
Vinura Dhananjaya, Piyumal Demotte, Surangika Ranathunga and Sanath Jayasena . . . . . 7377

*Pre-training and Evaluating Transformer-based Language Models for Icelandic*
Jón Guðnason and Hrafn Loftsson . . . . . 7386