Francis F. Steen
University of California Los Angeles
Los Angeles, CA, USA

Anders Hougaard
University of Southern Denmark
Odense, Denmark

Jungseock Joo
University of California Los Angeles
Los Angeles, CA, USA

Inés Olza
University of Navarra
Pamplona, Spain

Cristóbal Pagán Cánovas
University of Navarra
Pamplona, Spain

Anna Pleshakova
University of Oxford
Oxford, United Kingdom

Soumya Ray
Case Western Reserve University
Cleveland, OH, USA

Peter Uhrig
FAU Erlangen-Nürnberg
Erlangen, Germany

Javier Valenzuela
University of Murcia
Murcia, Spain

Jacek Woźny
University of Wrocław
Wrocław, Poland

Mark Turner
Case Western Reserve University
Cleveland, OH, USA

Toward an infrastructure for data-driven multimodal communication research

https://doi.org/10.1515/lingvan-2017-0041

ABSTRACT

Research into the multimodal dimensions of human communication faces a set of distinctive methodological challenges. Collecting the datasets is resource-intensive, analysis often lacks peer validation, and the absence of shared datasets makes it difficult to develop standards. External validity is hampered by small datasets, yet large datasets are intractable. Red Hen Lab spearheads an international infrastructure for data-driven multimodal communication research, facilitating an integrated cross-disciplinary workflow. Linguists, communication scholars, statisticians, and computer scientists work together to develop research questions, annotate training sets, and develop pattern discovery and machine learning tools that handle vast collections of multimodal data, beyond the dreams of previous researchers. This infrastructure makes it possible for researchers at multiple sites to work in real-time in transdisciplinary teams. We review the vision, progress, and prospects of this research consortium.

Key words

multimodality; machine learning; automated parsing; corpora; research consortia

1. *Introduction*

Human face-to-face communication has always taken place across multiple modalities: through gesture, facial expression, posture, tone of voice, pacing, gaze direction, touch, and words. Elaborate multimodal communication is a central and constantly active part of human cognition, in science, technology, engineering, mathematics, art, religion, crafts, social interaction, learning, innovation, memory, attention, travel, and all other activities, whether goal-based or not. Cultures invest heavily to support this aspect of human life: classical cultures emphasized the importance of rhetorical training, and today's world is crowded with novel technologies of multimodal communication, from television to social media, creating an unprecedented trove of digital records. Communication skills involve higher-order cognition, precisely timed movements, delicately modulated sounds, conceiving of the mental states of others from moment to moment, dynamically coordinating with other agents, and a high level of contextual awareness (Duranti & Goodwin 1992; Clark 1996).

From Panini (Sharma 1987–2003) to Chomsky and McGilvray (2012), the systematic study of human communication has been largely focused on the written representation of language: understandably so, as it is highly structured and can be shared in the process of describing and explaining it. The full multimodal dimensions of communication have a very brief history of

scholarship and present new methodological challenges. Communicative behavior must be recorded with resource-intensive audio-visual technologies. Since the range of available expressions is so wide, individual researchers need to specialize in specific modalities and constructions. Naturalistic data are typically not readily available; boutique collections from lab recordings take their place. Large-scale datasets are required for systematic study, yet no single researcher has the required time, resources, or motivation to create them. Worse, any group of researchers that succeeds in generating a massive dataset of multimodal communication will quickly be overwhelmed, since linguistics lacks the tools for mechanically searching and characterizing the material.

For the study of digitized written language, a wide range of technologies and tools available has been developed over the past decades in the contexts of corpus linguistics, computational linguistics, and artificial intelligence. In this context, carefully sampled corpora such as the British National Corpus (BNC) as well as larger, less carefully sampled corpora have emerged along with corpus retrieval software such as BNCweb (Hoffmann & Evert 2006) and its generalized and extended version, CQPweb (Hardie 2012), or the com- mercial Sketchengine (www.sketchengine.co.uk). While these are outstanding examples of research at the intersection of computer science and linguistics, they have not yet embraced the full multimodal spectrum of human communication, creating a well-defined disciplinary and interdisciplinary challenge.

At the same time, the social landscape of communication has exploded with new multimodal technologies, from television to social media, intruding on our most personal as well as our most public communicative functions. Since these exchanges are taking place in digital form, the age-old problem of how to capture multimodal communication in a naturalistic setting is now fully tractable. This drastically reduces the costs incurred by capturing and transcribing naturalistic spoken language such as the audio recordings collected for the spoken demographic section of the BNC (see Crowdy 1995 for details). So far, only modest attempts have been made to add audiovisual data to existing corpora; for instance, the Russian National Corpus (www.ruscorpora.ru) began in 2010 to include a selection of recordings and movies from 1930 to 2007, with 4.6 million words of transcripts. Only recently has the construction of massive multimodal corpora become feasible.

Such large-scale datasets present both an opportunity and a challenge for linguists. On the one hand, we can now attest the presence, context, and frequency of known constructions in ecologically valid datasets, extending, correcting, and validating decades of laboratory research. On the other hand, these datasets are so large that they quickly swamp manual analysis. The

challenge must be met with a new level of interdisciplinary collaboration between linguists and computer scientists. Both fields have much to gain. Computational researchers gain insights into natural modes of communication, useful for designing good user interfaces and natural interactions with robotic systems; linguists gain the knowledge of tools and methods from computer vision, audio signal processing and machine learning to analyze large amounts of data. We see an opportunity to create a collaborative and distributed social and physical infrastructure for data-driven multimodal communication research.

We can draw inspiration from other sciences in which cooperatives of researchers with diverse backgrounds were established to share the gathering of data and the development of tools and analysis in real time. Faced with a similar mix of massive new datasets and a demand for radically new methodologies, astronomy and genetics have undergone a comparable transformation in their disciplinary practices and thrived. Genomics dramatically speeded up its advances by creating worldwide consortia of researchers using collaborative web platforms; see for instance the Mission Statement and Framework of the Global Alliance for Genomics and Health (www.genomicsandhealth.org). In neuroscience, there are analogous initiatives, such as the Brainhack project (Craddock et al. 2016).

Such cooperative frameworks exhibit novel social dynamics and facilitate rapid disciplinary progress. On the one hand, information begins to flow across disciplinary boundaries, giving computational researchers access to novel real-world problems and researchers in the target domain exposure to new methods and skills. See, for example, the recent applications of multimodal computational methods in political science (Joo et al. 2015), psychology (Martinez 2017), or cognitive film studies (Suchan & Bhatt 2016). Just as important, the research results of one researcher — the actual data selection, annotation, and analysis — can be provided immediately and substantively to the whole community, and replicated and built on in a meaningful way.

In this paper, we describe the Distributed Little Red Hen Lab, a global laboratory and consortium designed to facilitate large-scale collaborative research into multimodal communication. As part of this project, we collect data on multimodal communication on a large scale, provide computational and storage tools to manage data and aid in knowledge discovery, and provide means of iterative improvement by integrating the results and feedback of researchers into the project.

Red Hen's vision and program arise naturally from considerations that are common and frequent across all the sciences, concerning how to improve the way we do science — by developing an extensive and constantly-networked cooperative, by developing sociological patterns of extensive

real-time collaboration across the cooperative, and by aggregating big data and developing new methods and tools that are deployed across the cooperative. Such considerations have become inescapable for several disciplines, from biology to materials science, linguistics to archeology, genomics to neuroscience, astronomy to computer science. Red Hen brings these impulses to the science of human multimodal communication. As an example, for the past 4 years, Red Hen has partnered with Google Summer of Code to connect Computer Science students from around the world with expert mentors, generating a suite of new tools to analyze human communication.

Red Hen is not designed to be a service. Instead, it provides a framework for collaboration, pooling expertise and resources. Access to the Red Hen tools and data are provided through the project website (www.redhenlab.org), where researchers can both access data and contribute or provide feedback to the Red Hen project.

A core activity of Red Hen Lab is an international effort to create the physical and social infrastructure needed for the systematic study of multimodal communication. Key elements are data collection, data mining tool development, and search engines.

2. *Generating massive multimodal datasets*

A shared dataset is an essential aspect of the infrastructure required for data-driven multimodal communication research. Red Hen is open to datasets in any area in which there are records of human communication. This includes text, speech and audio recordings in any language, infant vocalization, art and sculpture, writing and notation systems, audiovisual records, architecture, signage, and of course, modern digital media. Records and methods related to nonhuman communication or communication between species (e.g. border collies responding to pointing gestures) are also naturally of interest to Red Hen. In principle, any recording in any format of any human communication is suitable for inclusion in the archive, which consists of networked data across the Red Hen cooperative, either natively digital or converted to digital form.

The most efficient way to acquire a massive multimodal and multilingual dataset is to record television, a task that can be fully automated. Fortunately, section 108 of the U.S. Copyright Act authorizes libraries and archives to record and store any broadcast of any audiovisual news program and to loan those data, within some limits of due diligence for the purpose of research. The NewsScape Archive of International Television News (www.newsscape.library.ucla.edu) is Red Hen's largest; as of November 2017, it included broadcasts from 51 networks, totaling 350,000 hours and occupying 120 terabytes. The collection dates back to 2005 and is growing at around

5,000 shows a month. It is an official archive of the University of California, Los Angeles (UCLA) Library, the digital continuation of UCLA's Communication Studies Archive, initiated by Paul Rosenthal in 1972. The analog collection is in the process of being digitized, promising to add additional years of historical depth to the collection. Under Red Hen, it has been expanded to record television news in multiple countries around the world, curated by local linguists participating in the Red Hen project. The NewsScape Archive now includes, in rough order of representation, broadcasts in English, Spanish, German, French, Norwegian, Swedish, Danish, Continental Portuguese, Brazilian Portuguese, Russian, Polish, Czech, Flemish, Persian, Italian, Arabic, and Chinese. The system is fully automated and scales easily, using credit-card-sized Raspberry Pi capture stations running custom open-source software.

This television news dataset includes hard and soft news, including talk shows and comedy, along with B-roll of surveillance video, crowd-sourced videos, recordings of public events where participants do not even know they are being recorded, etc. These genres contain a range of registers that include banter, unscripted conversations, and playful interviews. The studio components of the television news shows in NewsScape typically also contain a great amount of unscripted or partially improvised communicative events. The most constrained register, in which a speaker reads a text or recites a pre-prepared speech more or less verbatim, is no longer the standard way to communicate on television. This makes NewsScape a rich resource for studying largely spontaneous or unconscious aspects of multimodal communication, along with communicative behaviors associated with a range of formal registers.

Red Hen's infrastructure and tools also permit the incorporation of existing datasets, such as handcrafted collections of experimental data. The news material constitutes the bulk of the current collection, as this content is clearly protected by the US Copyright Act. Smaller datasets generated by individual researchers and teams, including student projects, are being added and described, and will be the subject of future publications.

Red Hen proposes two complementary strategies to deal with recordings that are protected by confidentiality laws. First, although lab recordings are typically protected by Institutional Review Board regulations, it seems plausible that an IRB might approve machine analysis of such recordings. Results of such analysis may be shareable, provided the data is anonymized. Second, video recordings can be submitted to a sketch filter, which removes textures critical to personal identification, yet retains structural elements of multimodal communication (Diemer et al. 2016).

3. Creating and searching metadata and annotations

Vast multimodal datasets are a boon and a curse. Linguists need them to validate existing constructions in ecologically valid datasets, and they can also revel in the prospect of testing an entire generation of new hypotheses, asking questions we simply lacked the data to answer. However, to effectively convert such data to knowledge, we need automated search capabilities, and to search, we need machine-readable transcripts, ideally enriched with metadata and annotations. Red Hen's annotation process relies on a multi-level feedback process between linguists and computer scientists, aimed at training computers to perform tasks that generate annotations according to the linguist's specifications.

The video stream is compressed to a 640 × 480 or similar picture size at 450 kbps; the audio stream is a stereo signal with a sampling rate of 44.1 kHz compressed to a bitrate of 96 kb/s. Red Hen expects that the 44.1 kHz sampling rate and the 96 kb/s bitrate will be sufficient to make most of the audible frequencies usable for spectrograms, but detailed tests are yet to be conducted.

Red Hen textual data is encoded in UTF-8, using the universal standard of comma-separated values, and named to identify the time, source, and type of the recording. The data is stored on UCLA Library servers and elsewhere within the Red Hen network as needed. This provides the input to a variety of custom search engines.

A series of pipelines process these data, using customized open-source software. Some tools require relatively little customization and can be deployed without deep modifications. For example, transcripts are automatically extracted from television video in the form of subtitles. For US broadcasts, commercials are automatically detected and annotated. Additional text that is written on the television screen is also extracted, using tesseract-ocr (www.github.com/tesseract-ocr) with significant customizations for eight different languages, examining frames at one-second intervals and retaining screen placement information.

In multimodal data analysis, timing is of the essence. Centisecond timestamps in UTC permit precise correlations of data extracted from different modalities. To validate a multimodal construction, we need reliable timestamps at the word level, so that individual words can be shown to co-occur with a gesture of the eyes, the face, the shoulders, the arms, and the hands. To achieve this, the caption text is first parsed into sentences, using custom software developed by manual inspection of abbreviations and conventions characteristic of the medium. These sentences are fed into Stanford CoreNLP (www.stanfordnlp.github.io/CoreNLP), a set of natural language processing utilities providing parts of speech, lemmas, and named entities in half a

dozen languages. Because television captions are typically created on the fly by professional captioners, they lag behind the speech and video stream by a low but variable number of seconds. Red Hen uses the open-source Gentle project (www.lowerquality.com/gentle) to align the text with the audio, generating precise timestamps for each word. The percentage of words that Gentle succeeds in matching to the audio stream also gives us a rough measure of the quality of the transcript.

Red Hen also provides an enormous dataset annotated for frames using the FrameNet project's (www.framenet.icsi.berkeley.edu) annotation scheme. To our knowledge, it is the largest dataset so annotated. The SEMAFOR project (www.ark.cs.cmu.edu/SEMAFOR) performs an automatic analysis of the frame-semantic structure of English text, using the FrameNet 1.5 release. Frame names, frame elements, and semantic role labeling results are available for around 200 million sentences.

The metadata and annotations, along with the video and audio, can be accessed by Red Hen members through the Edge search engine (available via www.newsscape.library.ucla.edu), which provides an easy and user- friendly web-based user interface. Linguists may prefer the CQPweb search engine, which provides access to syntactic categories and has the full support of its configurable query language.

A hallmark of Red Hen is to put the computational tools into streamlined production pipelines on high-performance computing clusters, so that the results are available in search engines for all users. There are significant advantages of scale to this approach, in that metadata extraction can be automated, plowing through hundreds of thousands of files at a very low marginal cost per file. The shared dataset and computational resources available to the Red Hen community lower the barriers for deploying, customizing, and developing new computational tools. The shared infrastructure model also means that a single individual or team's contribution generates benefits for a large group of people: each person's quantum of effort is multiplied by scale and automation. This dynamic dramatically lowers the cost for everyone of accessing the potential benefits of such tools, without requiring a large personal investment in mastering their logic and operation, thus spurring the discipline of linguistics forwards.

4. Automated multimodal search

Off-the-shelf tools, however, are not always available for the research Red Hen wants to conduct. The tools described above enable a variety of detailed textual searches through Red Hen's massive dataset. For many linguistic tasks, it is valuable to be able to search visually rather than textually through

the data. For example, for gesture research, we might envision executing a search of visual features through available videos to find instances of a particular gesture. Red Hen provides a search interface (Figure 1) aligned to this need, developed collaboratively by linguists and computer scientists on our team, an example of the kind of interdisciplinary collaboration common in Red Hen.

The key elements are a set of "detectors" that detect various visual aspects of a scene. In the current version, these detectors are focused on detecting fragments of gestures such as hand, head and lip motion. Red Hen envisages that, as the system is used and new tasks are proposed, additional detectors will be incorporated. Some of these detectors make use of machine learning models that are learned from data using supervised or unsupervised learning methods. The design of the system has three goals in mind. First, all detectors are combined into a single unified system. Second, it is easy to add or remove detectors. Third, detectors can be updated as human feedback is obtained. With a video file as an input, we perform the following actions:

1. Detect faces for each frame
2. Detect hands for each person in each frame
3. Detect head, hand, shoulder, lip motion for each person in each frame
4. Detect (a subset of) timeline gestures in each frame where hand motion is present.

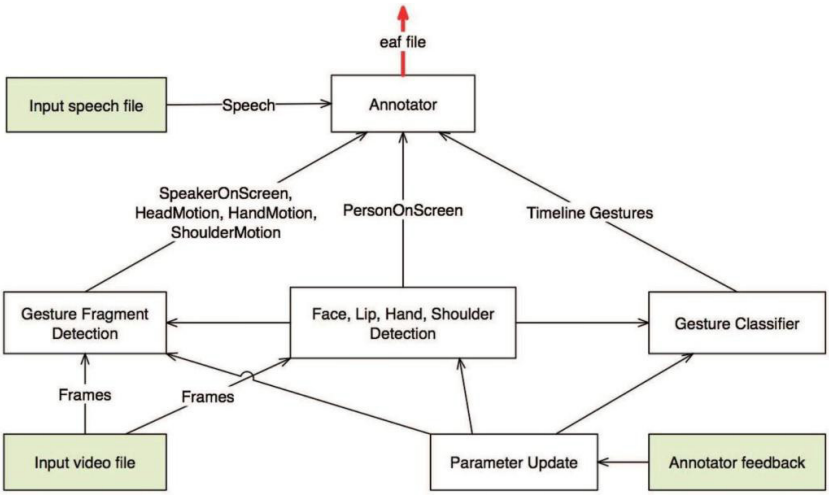

Figure 1. Architecture for integrated automatic feature detection, expert feedback (green box, lower left) and classifier enhancement in the visual search engine integrated with Red Hen (see text). Green boxes represent inputs. The others are modules in the search engine.

For a detailed discussion of the algorithms involved, see Turchyn et al. (in press). The output of this system is obtained in two ways: (i) as filters in CQPWeb and (ii) as annotations in ELAN annotation files. ELAN (https://tla.mpi.nl/tools/tla-tools/elan/) is a popular open-source application developed by the Max Planck Institute for Psycholinguistics to enable end users to perform manual tagging on audiovisual or text files, which are stored

in a custom XML format with the extension "eaf." In eaf files, annotations are arranged into tiers. For example, for an annotation file associated with a video, a "PersonOnScreen" tier is added, which marks all frames where a person was on the screen. A sub-tier of this is "SpeakerOnScreen," where the person was also speaking. The interface of ELAN then allows a user to jump between these annotations, so that, for example, when looking for specific gestures, all parts of the video where there were no people on screen can be easily skipped, simplifying the annotation process. Tiers are arranged in a coarse-to-fine manner.

The learned detectors are imperfect, and the fine detectors that are looking for small body part motions are more imperfect than the coarse detectors. Imperfections can be both false positives (a fragment is detected where there is none) or false negatives (no fragment is detected although it is present). Imperfections arise because the structure of frames in the videos in Red Hen is very complex, so that it is often extremely difficult to detect precisely small motions or parts of the body. We hypothesize that these imperfections will be reduced as more detectors are added and the existing ones updated. In the meantime, the system is still useful as a way to reduce effort during the annotation process, especially for tasks where some examples of specific elements such as gestures are desired. On the other hand, if the task involves a scenario such as counting the frequency of a certain gesture, then the coarse tiers could be used to rule out irrelevant parts of a video. However, a significant part of the remainder may need to be manually inspected to avoid the possibility that the fine detectors miss a region of interest.

An important part of our system is the ability of the detectors to update themselves as data becomes available in the form of human annotations. For example, if someone were to annotate some frames in a video as a timeline gesture, that detector could use the annotation to improve its ability to detect other timeline gestures. To allow this, every detector has a set of tunable parameters. Every so often, the system will collect all annotations relevant to a detection and update the values of these parameters so that the accuracy of a detector at its task improves. This is done in a transparent manner in the back end; once updated, a new set of detectors can be applied to the data, creating new results that flow automatically to the front end, without need of interaction from the user.

5. *Red Hen tools for linguistic research*

The availability of tagged and searchable multimodal big data opens up new opportunities for linguistic research, extending the utility of large corpora noted by Davies (2015). A central topic is the detection and characterization

of multimodal constructions. Nesset et al. (2013) describe how to use text search in a massive dataset to locate video clips of multimodal constructions; the text search was run on both English and Russian. The goal was to analyze what happens to Russian and English deictics in the context of a television news story, investigating the use of "the five Russian deictic words that correspond to the English meanings 'here' and 'now': zdes', tut, sejčas, teper' and vot." These are forms with routine if quite complicated use in actual scenes of classic joint attention. They acquire slightly different sets of distributions for TV news, where they have distinct radial category profiles, in the sense that they display different centers of gravity in the semantic network. The authors propose the "Minimal Adaptation Hypothesis," "according to which language makes adaptations that are as small as possible when applied to a new setting, such as the one created by TV." Viewers typically do not recognize consciously the adjustments in the meaning of *here*, *now*, *zdes'*, *tut*, *sejčas*, *teper'* and *vot* when they are deployed in such settings. The adjustments are considerable and patterned, but typically go unnoticed. Turner (2017a, 2017b) builds on this analysis by analyzing Red Hen data including instructional videos and films. Joo et al. (2017) outline and exemplify how machine learning can be recruited to automate tagging of individual expressive multimodal constructions, such as side-eye — a quick glance to the side employed communicatively. Once they are tagged automatically, they can be located in the big dataset by the linguist. Steen and Turner (2013) show how Red Hen can be used to bring powerful new computational tools to a classic method of linguistics: the analysis of constructions. They ran computational text search on 2.5 billion words to locate configurations of related linguistic constructions, and to export them to a csv file for analysis by inspection or by using the R statistics software package, which only takes minutes on a modern server. The network of constructions chosen for this demonstration consisted of the basic XYZ construction ("Causation is the cement of the universe") and its related constructions, as discussed in the article. They showed that, in this way, Red Hen can provide data very rapidly on constructions even if they are quite infrequent in discourse. They also showed how Red Hen can be used to search for patterns of blending of constructions that result in "mistakes" of the sort that are routinely edited out of printed texts. These are but a few examples of the way in which a vast number of traditional hypotheses in linguistics can be tested against tagged and searchable big multimodal data. For other examples, see Hoffmann (2017), Li et al. (2017), Turner (2015, 2017a, 2017b), Zima (2014a, 2014b, 2017), Pagán Cánovas & Valenzuela (2017).

Any new project wishing to explore a large dataset can benefit from the computational tools developed for NewsScape, and perhaps more im-

portantly from the expertise of a community of researchers who have used Red Hen resources in their own research projects.

6. Conclusion

Advances in technology on multiple fronts has brought the full spectrum of human communication within reach of systematic study, promising to revolutionize our understanding of linguistics and communication. Red Hen Lab coordinates a multi-disciplinary effort to realize this promise. It provides the technical means and the social networks to collect and store vast datasets, and institutes an integrated workflow uniting linguists, communication scholars, statisticians, and computer scientists in joint research projects. This workflow leverages the discriminating judgment of the individual researcher and the annotations of his or her students into training data for new applications of machine learning and the development of automated computational tools running on high-performance computing clusters. By combining human feedback on the output of these new tools, Red Hen is able to generate incremental improvements in identifying promising patterns and constructions in multimodal communication. New findings are not only published, but integrated into the shared dataset, facilitating rapid cumulative progress. The collaboration between hitherto isolated research areas that Red Hen facilitates is a spur to new research questions and enables novel problems and techniques to emerge.

Acknowledgments

Red Hen Lab acknowledges the support of the National Science Foundation CNS 1028381 and 1027965, the Anneliese Maier Research Prize from the Alexander von Humboldt Foundation, the Deutsche Forschungsgemeinschaft, the Research Council of Norway, KONWIHR, Google Summer of Code, UCLA Library, UCLA IDRE, and Case Western Reserve University UTech, most notably its High Performance Computing Group.

References

Chomsky, Noam, James McGilvray 2012: *The science of Language*. Cambridge: Cambridge University Press.

Clark, Herbert H. 1996: *Using Language*. Cambridge: Cambridge University Press.

Craddock, R. Cameron, Daniel S. Margulies, Pierre Bellec, B. Nolan Nichols, Sarael Alcauter, Fernando A. Barrios, Yves Burnod, Christopher J. Cannistraci, Julien Cohen-Adad, Benjamin De Leener, Sebastien Dery, Jonathan Downar, Katharine Dunlop, Alexandre R. Franco, Caroline

Seligman Froehlich, Andrew J. Gerber, Satrajit S. Ghosh, Thomas J. Grabowski, Sean Hill, Anibal Sólon Heinsfeld, R. Matthew Hutchison, Prantik Kundu, Angela R. Laird, Sook-Lei Liew, Daniel J. Lurie, Donald G. McLaren, Felipe Meneguzzi, Maarten Mennes, Salma Mesmoudi, David O'Connor, Erick H. Pasaye, Scott Peltier, Jean-Baptiste Poline, Gautam Prasad, Ramon Fraga Pereira, Pierre-Olivier Quirion, Ariel Rokem, Ziad S. Saad, Yonggang Shi, Stephen C. Strother, Roberto Toro, Lucina Q. Uddin, John D. Van Horn, John W. Van Meter, Robert C. Welsh, Ting Xu 2016: Brainhack: A collaborative workshop for the open neuroscience community. *GigaScience* 5.16. DOI 10.1186/s13742-016-0121-x.

Crowdy, Steve 1995: The BNC spoken corpus. In: Geoffrey N. Leech, Greg Myers, Jenny Thomas (eds.) 1995: *Spoken English on Computer: Transcription, Mark-up and Application*. Harlow: Longman, 224–235.

Davies, Mark 2015: The importance of robust corpora in providing more realistic descriptions of variation in English grammar. *Linguistics Vanguard* 1.1, 305–312.

Diemer, Stefan, Marie-Louise Brunner, Selina Schmidt 2016: Compiling computer-mediated spoken language corpora: Key issues and recommendations. *International Journal of Corpus Linguistics* 21.3, 349–371.

Duranti, Alessandro, Charles Goodwin (eds.) 1992: *Rethinking Context: Language as an Interactive Phenomenon*, vol. 11. Cambridge: Cambridge University Press.

Hardie, Andrew 2012: CQPweb — Combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics* 17.3, 380–409.

Hoffmann, Sebastian, Stefan Evert 2006: BNCweb (CQP Edition) — The marriage of two corpus tools. In: Sabine Braun, Kurt Kohn, Joybrato Mukherjee (eds.) 2006: *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*. Frankfurt am Main: Peter Lang, 177–195.

Hoffmann, Thomas 2017: Multimodal constructs — Multimodal constructions? The role of constructions in the working memory. *Linguistics Vanguard* 3.s1. DOI 10.1515/lingvan-2016-0042.

Joo, Jungseock, Francis F. Steen, Song-Chun Zhu 2015: Automated facial trait judgment and election outcome prediction: Social dimensions of face. *Proceedings of the IEEE International Conference on Computer Vision*. Los Alamitos, CA: Institute of Electrical and Electronics Engineers, 3712–3720.

Joo, Jungseock, Francis F. Steen, Mark Turner 2017: Red Hen Lab: Dataset and tools for multimodal human communication research. In: Mehul Bhatt, Kristian Kersting (eds.) 2017: *KI - Künstliche Intelligenz. Special Edition*. Berlin Heidelberg: Springer, 1–5.

Li, Weixin, Jungseock Joo, Hang Qi, Song-Chun Zhu 2017: Joint image-text

news topic detection and tracking by multimodal topic and-or graph. *IEEE Transactions on Multimedia* 19.2, 367–381.

Martinez, Aleix M. 2017: Computational models of face perception. *Current Directions in Psychological Science* 26.3, 263–269.

Nesset, T., A. Endresen, L. Janda, A. Makarova, F. Steen, M. Turner 2013: How 'here' and 'now' in Russian and English establish joint attention in TV news broadcasts. *Russian Linguistics* 37, 229–251. DOI 10.1007/s11185-013-9114-x.

Pagán Cánovas, Cristóbal, J. Valenzuela 2017: Timelines and multimodal constructions: Facing new challenges. *Linguistic Vanguard* 3.s1. DOI 10.1515/lingvan-2016-0087.

Sharma, Rama Nath 1987–2003: *The Astadhyayı of Panini*. 6 Vols. New Delhi: Munshiram Manoharlal.

Steen, Francis F., Mark Turner 2013: Multimodal construction grammar. In: Michael Borkent, Barbara Dancygier, Jennifer Hinnell (eds.) 2013: *Language and the Creative Mind*. Stanford, CA: CSLI Publications, University of Chicago Press, 255–274.

Suchan, Jakob, Mehul Bhatt 2016: The geometry of a scene: On deep semantics for visual perception driven cognitive film studies. *IEEE Winter Conference on Applications of Computer Vision (WACV)*.Piscataway, NJ: Institute of Electrical and Electronics Engineers, 1–9.

Turchyn, Sergiy, Inés Olza Moreno, Cristóbal Pagán Cánovas, Francis F. Steen, Mark Turner, Javier Valenzuela, Soumya Ray (in press): Gesture annotation with a visual search engine for multimodal communication research. *Proceedings of the Thirtieth Annual Conference on Innovative Applications of Artificial Intelligence (IAAI-18)*.

Turner, Mark 2015: Blending in language and communication. In: Ewa Dąbrowska, Dagmar Divjak (eds.) 2015: *Handbook of Cognitive Linguistics*. Berlin: De Gruyter Mouton, 211-232.

Turner, Mark 2017a: Multimodal form-meaning pairs for blended classic joint attention. *Linguistics Vanguard* 3(s1). DOI 10.1515/lingvan-2016-0043.

Turner, Mark 2017b: Polytropos and communication in the wild. In: Barbara Dancygier (ed.) 2017: *The Cambridge Handbook of Cognitive Linguistics*. Cambridge: Cambridge University Press, 93–98.

Zima, Elisabeth 2014a: English multimodal motion constructions: A construction grammar perspective. *Studies van de BKL - Travaux du CBL - Papers of the LSB*, Volume 8. URL: www.uahost.uantwerpen.be/linguist/SBKL/sbkl2013/Zim2013.pdf. ED: 1 Dec. 2017.

Zima, Elisabeth 2014b: Gibt es multimodale Konstruktionen? Eine Studie zu [V(motion) in circles] und [all the way from XPREPY].*Gesprächsforschung-Online Zeitschrift zur verbalen Interaktion* 15.1–48. URL: www.gespraechsforschung-ozs.de/fileadmin/dateien/heft2014/ga-zima.pdf. ED: 1 Dec. 2017.

Zima, Elisabeth 2017: Multimodal constructional resemblance: The case of English circular motion constructions. In: Francisco Ruiz de Mendoza Ibáñez, Alba Luzondo, Paula Pérez-Sobrino (eds.) 2017: *Constructing families of constructions*. Human Cognitive Processing Series. Amsterdam: John Benjamins, 301-337.

Streszczenie

W stronę infrastruktury badań komunikacji multimodalnej na wielkich zbiorach danych

Celem artykułu jest wyjaśnienie, w jaki sposób Red Hen Lab tworzy duże multimodalne bazy danych, jak powstają metadane oraz zautomatyzowane narzędzia umożliwiające precyzyjne wyszukiwanie oparte na parametrach zarówno tekstowych, jak i graficznych.

Komunikacja między ludźmi zawsze była i jest multimodalna. Porozumiewamy się za pomocą gestu, wyrazu twarzy, postawy, tonu głosu, kierunku spojrzenia, dotyku i słowa. Historia postępów rozumienia i opisu cech ludzkiej komunikacji ma już ponad dwa tysiące lat (za ich początek często przyjmuje się Ośmioksiąg Paniniego z V-IV w p.n.e.), jednak – głównie ze względów technicznych – badania te były oparte w przeważającej mierze na słowie pisanym. Szybki rozwój technologii komunikacyjnych ostatnich dekad – zwłaszcza cyfrowy zapis obrazu i dźwięku, znaczny wzrost prędkości przesyłania danych oraz wzrost pojemności nośników elektronicznych – pozwala na lepsze zrozumienie rzeczywistej złożoności komunikacji między ludźmi. To właśnie ten cel przyświecał założycielom konsorcjum naukowego Distributed Little Red Hen Lab (www.redhenlab.org) - Markowi Turnerowi (CWRU) i Francisowi Steenowi (UCLA). Konsorcjum to stanowi globalną wspólnotę badawczą skupiającą językoznawców i informatyków reprezentujących około 30 instytucji naukowych z całego świata (między innymi Uniwersytetu Oxfordzkiego, Uniwersytetu Kalifornijskiego w Los Angeles, Uniwersytetu Alberty, Uniwersytetu Wrocławskiego).

Celem Red Hen Lab są badania komunikacji multimodalnej, a zwłaszcza tworzenie infrastruktury do takich badań. Częścią tej infrastruktury jest na przykład największy na świecie, stale rozbudowywany korpus multimodalny (który łączy dane audio, wideo oraz tekst) – UCLA NewsScape Library – umożliwiający wyszukiwanie zarówno tekstowe, jak i graficzne (np. wyszukiwanie gestów lub innych elementów języka ciała towarzyszących mowie). W roku 2018 korpus ten zawiera (nie licząc metadanych) już ponad cztery miliardy słów (a więc 40 razy więcej niż np. korpus BNC) i kilkaset tysięcy godzin nagrań wideo. Co miesiąc zasoby UCLA NewsScape Library powiększają się o około 5000 programów nagrywanych z 51 stacji telewizyjnych nadających w języku angielskim, hiszpańskim, niemieckim, francuskim, norweskim, szwedzkim, duńskim, portugalskim (dialekt kontynentalny i brazylijski), rosyjskim, polskim, czeskim, flamandzkim, perskim, włoskim, arabskim i chińskim. Programy nagrywane są z anten telewizji naziemnej przez rozmieszczone na całym świecie stacje nagrywające (capture stations). Podstawowym wyposażeniem takiej stacji jest komputer Raspberry Pi (wielkości karty kredytowej) oraz konwerter przetwarzający sygnał analogowy z anteny naziemnej na sygnał cyfrowy (najczęściej jest to urządzenie HDHomeRun firmy SiliconDust). Raspberry Pi jest w pełni funkcjonalnym komputerem PC. Jego zaletą oprócz małych rozmiarów jest niskie zużycie energii (poniżej 1W), co ma duże znaczenie, ponieważ stacja nagrywająca pracuje w trybie ciągłym. Oprogramowanie stacji nagrywającej stanowi pakiet programów z upublicznionym kodem źródłowym, pozwalający na praktycznie bezobsługową pracę. Każda stacja nagrywająca ma jednak swojego lokalnego opiekuna współpracującego z Red Hen Lab. Zadaniem opiekuna stacji jest wybór programów do nagrywania,

okresowa kontrola poprawności pracy stacji i zgłaszanie ewentualnych usterek. Jednym z problemów związanych z codziennym nagrywaniem wybranych programów jest zmieniająca się ramówka telewizyjna. Początkowo do zadań opiekuna stacji należało śledzenie programu telewizyjnego i ręczna korekta godzin nagrywania. Jest to jednak system zbyt zawodny. Dlatego jedna ze stacji Red Hen Lab codziennie pobiera program telewizyjny z lokalnych stron internetowych na całym świecie i po przetworzeniu rozsyła zaktualizowane dane do poszczególnych stacji. Innym istotnym problemem jest zmieniająca się jakość transmisji naziemnej (czasem uzależniona nawet od pogody). Oprogramowanie stacji pozwala na stałe kontrolowanie siły i jakości sygnału oraz jakości samego nagrania. Przy niewielkiej ilości błędów stacja podejmuje próbę ich automatycznego naprawienia. Jeśli mimo to nagranie jest nadal niskiej jakości, zostaje ono usunięte, a opiekun stacji otrzymuje automatyczne powiadomienie o problemach z nagrywaniem. Bardzo ważną częścią nagrania są napisy, najczęściej pobierane z odpowiedniej strony teletekstu (jeśli taka istnieje) lub wyodrębniane z plików wideo (mpeg) przy pomocy odpowiedniego oprogramowania. Jakość napisów jest także kontrolowana na bieżąco przez stację nagrywającą. Oprogramowanie stacji umożliwia także automatyczne usuwanie błędów w plikach z napisami (na przykład nieciągłości czasowych). Pliki tekstowe Red Hen Lab są kodowane w UTF-8 w uniwersalnym formacie CSV (comma-separated values) i zawierają informację o czasie, źródle i typie nagrania. Pliki tekstowe oraz pliki wideo (po konwersji do rozdzielczości 640x480, 450 kbps, audio 44.1 kHz, bitrate 96 kb/s) przesyłane są do serwerów UCLA NewsScape Library.

Przesłanie wstępnie przygotowanych plików do serwerów UCLA nie kończy jeszcze procesu ich przetwarzania. Będą one teraz podlegać dalszej obróbce na kilku „liniach produkcyjnych" (pipelines) przygotowanych dla Red Hen Lab na podstawie oprogramowania z otwartym dostępem do kodu źródłowego. Niektóre elementy gotowego oprogramowania wymagały tylko niewielkich przeróbek, inne – istotnych zmian w celu dostosowania ich do potrzeb Red Hen Lab. Na przykład tesseract-ocr (github.com/ tesseract-ocr) dzieki dostosowaniu pozwala teraz na rozpoznanie tekstów pojawiających się na ekranie (jako część grafiki ekranowej) w 8 językach z częstością odświeżania 1 ekran/s. W analizie danych multimodalnych korelacja czasowa ma bardzo istotne znaczenie. Precyzyjna koordynacja danych pochodzących z różnych modalności jest możliwa dzięki zawartym w plikach tekstowych znacznikom czasu (timestamps) w UTC, które podają czas wyświetlania napisów na ekranie z dokładnością do setnej sekundy. Do dokładniejszej analizy, np. do ustalenia, które słowo współwystępuje z określonym gestem, potrzebne są jednak precyzyjniejsze dane, czyli znaczniki czasu osobno dla każdego słowa, ponieważ bardzo często zdarza się że napisy są opóźnione w stosunku do strumienia audio. Aby to osiągnąć, tekst napisów jest najpierw dzielony na pojedyncze zdania za pomocą programu Stanford CoreNLP (stanfordnlp.github.io/CoreNLP), który umożliwia analizę syntaktyczną w sześciu językach. Następnie wykorzystywany jest pakiet oprogramowania Gentle Project (lowerquality.com/gentle), który pozwala uzyskać znaczniki czasowe dla każdego słowa. Innym przykładem przetwarzania danych gromadzonych przez stacje nagrywające jest największy na świecie zbiór danych (około 200 milionów zdań) przygotowany przy pomocy oprogramowania The SEMAFOR project (http://www.ark.cs.cmu.edu/SEMAFOR) pozwalającego na annotację tekstu ze względu na ramy czasownikowe i role semantyczne zgodnie ze schematem The FrameNet Project (framenet.icsi.berkeley.edu). Pliki tekstowe oraz audiowizualne wraz z metadanymi i annotacjami są dostępne dla współpracowników Red Hen Lab przez prostą w obsłudze wyszukiwarkę Edge Search Engine (newsscape.library.ucla.edu).

Red Hen Lab umożliwia koordynację pracy wielu multidyscyplinarnych zespołów badawczych, które dzielą się wynikami swojej pracy. Jest to ten sam sieciowy model współpracy, który został wykorzystany np. przez Human Genome Project (genomicsandhealth.org). W obu przypadkach przed badaczami stoją podobne wyzwania - ogromne ilości danych wymagające tworzenia nowych narzędzi i nowych metod badawczych. Czy Sieci badawcze, takie jak Red Hen Lab (redhenlab.org), mogą stanowić skuteczną odpowiedź na te wyzwania? Zapraszam do przeczytania całego artykułu.

Streszczenie przygotował Jacek Woźny