Studying Time Conceptualisation via Speech, Prosody, and Hand Gesture: Interweaving Manual and Computational Methods of Analysis

Peter Uhrig[1,4], *Elinor Payne*[2], *Irina Pavlova*[2], *Ilya Burenko*[1,4], *Nathan Dykes*[3], *Mary Baltazani*[2], *Evie Burrows*[2], *Scott Hale*[2], *Philip Torr*[2], *Anna Wilson*[2]

[1]Technische Universität Dresden, Germany
[2]University of Oxford, United Kingdom
[3]Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
[4]Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI), Dresden/Leipzig
peter.uhrig@tu-dresden.de, anna.wilson@area.ox.ac.uk

Abstract

This paper presents a new interdisciplinary methodology for the analysis of future conceptualisations in big messy media data. More specifically, it focuses on the depictions of post-Covid futures by RT during the pandemic, i.e. on data which are of interest not just from the perspective of academic research but also of policy engagement. The methodology has been developed to support the scaling up of fine-grained data-driven analysis of discourse utterances larger than individual lexical units which are centred around 'will' + the infinitive. It relies on the true integration of manual analytical and computational methods and tools in researching three modalities – textual, prosodic[1], and gestural. The paper describes the process of building a computational infrastructure for the collection and processing of video data, which aims to empower the manual analysis. It also shows how manual analysis can motivate the development of computational tools. The paper presents individual computational tools to demonstrate how the combination of human and machine approaches to analysis can reveal new manifestations of cohesion between gesture and prosody. To illustrate the latter, the paper shows how the boundaries of prosodic units can work to help determine the boundaries of gestural units for future conceptualisations.

Index Terms: multimodal data processing, computer vision, future conceptualization, temporal gestures, prosody, automation of analysis

1. Introduction

Multimodal representations of time conceptualisations have been of interest to linguists, psychologists, and anthropologists but still remain under-researched (see e.g. Cooperrider et al. 2014). We pose the following questions:

- How do speech – textual and audio modes – and gesture work together to 'verbalise' future depictions?
- How can a combination of human and machine approaches help us research multimodal communication about futures effectively?
- Does the computer-assisted analysis of three modalities – text, prosody, and gesture – allow us to engage more successfully with the bigger question of what gesture is?

To answer these research questions, we adopted an interdisciplinary approach to analysis, which was data-driven and exploratory. We went where data took us, not disregarding data that did not fit our hypotheses at the outset of the project. We took a step back from focusing on lexico-semantic units as they co-occur with prosodic features and individual gestures, to consider larger spoken discourse units and gesticulation as they contribute to time conceptualization at the semantic-syntactic level as discourse unfolds. We have been gradually scaling up our analysis to test the insights emerging from our manual analysis and annotation for future speech and gestural markers. As our analysis progressed and challenges emerged, we worked on the development, customisation, and integration of computational tools to automate and hence speed up specific parts of our multimodal corpus analysis. Our computer-assisted manual annotation and analysis of speech and gesture have also shed some light on the question of determining gesture boundaries, which we report as a case study.

The use and creation of manual and computational methods and tools evolved in an interwoven and interdependent fashion. As our manual analysis progressed, we understood more about which computational tools and methods were needed. As we developed or customised our computational tools and methods, we had a better appreciation of how we needed to adjust our manual annotation and develop and adapt our methods for manual multimodal analysis. We present our analytical and computational tools below one by one for clarity, but it has to be borne in mind that the genesis of the individual tools and the manual analysis happened in parallel and in constant exchange between the various disciplines involved, as illustrated by the following examples:

As we engaged in exploratory ELAN annotation and analysis for speech and gestural markers of futures, we were developing hand detection pipeline for the use in CQPweb (Hardie 2012; see Section 2.4) relying on OpenPose (Cao et al. 2018; see Section 2.3), collecting more video data and developing the overall computational infrastructure for data processing and retrieval.

[1] We divide what can be viewed as one 'speech' modality into two – textual and prosody – for the purposes of our work combining human and machine analyses.

The results of our manual analysis informed our corpus queries in CQPweb and our annotation in the Rapid Annotator, as well as the further development of the computer vision pipeline with regard to active speaker detection and biometric clustering (speaker detection and recognition).

The annotation results in the Rapid Annotator motivated a deeper and more fine-grained manual analysis of 47 videos. The observation of patterns to do with gestural axes, directions, and zones led to us to develop a new scheme for ELAN annotation which incorporated manual annotations for hand gesture and prosody and automatic annotations for gesture zones and time series (see Sections 2.3, 2.6, 2.7 and 3).

2. Methodology

2.1. Dataset

Our focus on future conceptualization is reflected in our choice of data. We collected all available episodes of the show *SophieCo Visionaries*, which was produced by the Russian state-funded international broadcaster RT (formerly *Russia Today*). This show has an explicit focus on the future and thus offers a much higher density of time expressions than most other TV formats. The show is produced in the form of an interview that the host, Sophie Shevardnadze, conducts with an expert on the topic of the respective episode, typically with her in the studio and the guest brought in remotely. During the heyday of the Covid-19 pandemic, episodes were recorded at her home, too. We downloaded the show episodes from YouTube with yt-dlp[1], a fork of the popular youtube-dl[2], which allowed us to obtain additional metadata on the video, the top-level comments, and – most importantly for this paper – the automatically-generated subtitles, which we use as transcript. The SophieCo Visionaries corpus consists of 99 videos and approximately 460,000 words. It is part of a larger corpus with all shows in which Sophie Shevardnadze occurs, which comprises 1.5 million tokens and spans a timeframe from 2008-2022 with a total of 439 videos. Relevant videos were identified in our full RT collection by searching the metadata JSON files for different spelling variants of *sophieco* and *sophie shevardnadze*.

2.2. Text processing

For each video, the text and timestamps for the beginning and end of each word are extracted from the automatically generated subtitle file and converted to CoNLL-U format. During this process, the text is tokenised with SoMaJo[3] (Proisl & Uhrig 2016). In the next step, the raw text is extracted and processed with a fine-tuned version of Alam et al.'s tool for punctuation restoration[4], which inserts commas, question marks, periods, exclamation marks and dashes and inserts sentence boundaries after relevant punctuation marks. The fine-tuning was done on the Brown Corpus family and serves to expand the inventory of possible punctuation marks. The resulting punctuated text is transferred back into CoNLL-U format and tagged with UDPipe[5] (Straka & Straková) using the english-ewt model. Finally, a verticalised text file is created in a format compatible with the CQPweb, which incorporates both the annotated tokens and relevant text-level metadata. The resulting .vrt file can be uploaded to CQPweb after adding gesture information from the computer vision analysis.

2.3. Audiovisual processing

In the manual analysis, it became clear that much time was spent sifting through irrelevant examples. To restrict the query results to instances of the show host speaking, we deployed an Active Speaker Detection pipeline (ASD) to detect the host for each given scene. For ASD we used TalkNet (Tao et al. 2021), which combines visual features extracted by a convolutional neural network from the input video and audio features obtained by another convolutional network from the Mel-frequency cepstrum via a cross-attention mechanism.

For the host detection task, we used biometric clustering. We were able to cluster vector representations of detected persons and expected that representations of the host would be close to each other compared to the representation of the faces of interviewees. We split each video into scenes[6], tracked persons on each individual scene, and calculated the mean face vector representation[7] for each tracked person. After that we used the HDBSCAN[8] clustering algorithm in order to cluster the obtained vector representations. In most cases we obtained from 2 to 4 clusters, where one cluster corresponds to the host, another corresponds to an interviewee and other clusters comprise either vector representation of the host shot almost from the back or representations of persons from background videos or noise.

For each scene we assigned a cluster to every tracked person and calculated a centroid of each cluster and ran the DBSCAN clustering algorithm on these centroids. The resulting clustering met our goals since we were able to distinguish between the host and an interviewee with very high reliability, as we were able to validate by inspecting members of each centroid cluster.

With the Active Speaker Detection we now know whether and where the speaker is on the screen, and with the biometric clustering, we know whether it is Sophie Shevardnadze. However, many shots of her – in particular in the videos filmed at her home at the beginning of the pandemic – do not show her hands, which prevents us from the analysis of co-speech gesture. Thus, to find instances of her with visible hands, we deploy OpenPose (Cao et al. 2018) to annotate body pose, which means that a set of so-called keypoints for every person detected on screen is created, and the results are stored as coordinates in a data structure that can then be processed further. OpenPose sports three detectors, viz. body pose, face keypoints and hand keypoints. The hand keypoints are dependent on the identification of a wrist, which in turn is dependent on the identification of an elbow via which it is connected to the rest of the body. Due to the motion blur in the videos during fast hand movements, hand keypoints are not detected as reliably as we would like, and the wrists from the set of body pose keypoints have proven to be much more robustly detected, even though their detection rate also drops during fast movements.

[1] https://github.com/yt-dlp/yt-dlp

[2] https://youtube-dl.org/

[3] https://github.com/tsproisl/SoMaJo

[4] https://github.com/xashru/punctuation-restoration

[5] https://github.com/ufal/udpipe

[6] http://scenedetect.com/en/latest/

[7] https://github.com/timesler/facenet-pytorch

[8] https://github.com/scikit-learn-contrib/hdbscan

The OpenPose keypoints can also be used in the visualization of a continuous time series (see e.g. Pouw & Trujillo 2021), which we used for vertical and horizontal hand position, as shown in the videos presented in Section 3.

In addition, several further measures were derived from the OpenPose keypoints of the show host. After smoothing and low-pass filtering to remove jitter, the horizontal and vertical position of the hands were determined according to a set of pre-defined zones. For this, we had to normalise the speaker's size, which we achieved by centering on the average position of the speaker's nose in a given scene and expressing all distances in units corresponding to the distance between that average nose position and the average position of the neck keypoint. We then automatically created two tiers in the ELAN files (see Section 3), one for vertical and one for horizontal zone. We also determined the pointing direction of the index finger and the thumb, but so far these have not yet been evaluated. Further derivations will follow.

2.4. CQPweb

Our dataset is too large for manual inspection, which means that specialized corpus-linguistic software had to be used to retrieve relevant sections from the shows. We opted for CQPweb (Hardie 2012) with custom visualizations and plugins (see Uhrig 2022) that allowed us to directly watch the videos from the concordance and download the concordance with added video links to be used with Rapid Annotator (see Section 2.5).

In preparation for use with CQPweb, the video annotations were merged into the vertical files containing the linguistically-annotated words with their timing information from the YouTube captions, which we used to decide to which word the time-indexed video annotations should be attached. Due to the recurrent misses of wrists by OpenPose during fast hand movements on very short words (such as cliticized *'ll*), we did not only annotate for words being uttered while wrists were visible, but we also introduced a second set of annotations that tell us whether the host's wrists were seen within two seconds before and two seconds after the target word, which improves recall (i.e. the detection of true positive cases) at the cost of lowering precision (i.e. it then also detects more false positive cases).

The attributes encoded in the corpus include for each word form a part of speech, the lemma, a coarse-grained word class, a set of morphological features, a binary feature for the biometric detection of the show host on screen, a binary feature saying whether the show host speaks this word, and a total of 8 features for the detection of the right wrist and the left wrist respectively, both during the utterance of only the word itself and with 2 seconds to either side, and for the wrists of the show hosts and of other people. For each of these attributes, we transposed the confidence score given by OpenPose for the corresponding keypoint to an integer between 0 and 100 so that the researcher can decide whether they want to maximise precision by choosing a higher confidence threshold, risking missing relevant examples, or whether they want to maximise recall by choosing a lower confidence threshold, risking an increase in wrong detections in their results.

The power of CQPweb as set up here (see Uhrig 2022 for more details) lies in its ability to seamlessly combine the various modalities in fast searches of large datasets, i.e. we can look for modal verbs followed by the base form of a verb uttered by the show host while at least one of her wrists is visible on the screen. This can save large amounts of manual labour, as the following examples illustrate for our latest corpus of all RT shows featuring Sophie Shevardnadze (439 videos, 1.5 million words):

Table 1: *CQPweb queries with numbers of hits for purely linguistic queries and queries with added restrictions on the computer vision annotation*

Query	Explanation	Hits
[lemma="will" & pos="MD"]	*will* as a modal verb lemma (includes *'ll* and *wo* in *won't*)	6,768
[lemma="will" & pos="MD" & so_detected="1"]	as above, but the show host is detected on the screen	3,445
[lemma="will" & pos = "MD" & so_speaks="1"][1]	as above, but the show host is speaking	1,519
[lemma="will" & pos = "MD" & so_speaks="1" & (int(so_rw_offset_0) > 20 | int(so_lw_offset_0) > 20)]	as above, but at least one wrist is detected with a confidence score of more than 20 during the utterance of the word	985
[lemma="will" & pos = "MD" & so_speaks="1" & (int(so_rw_offset_2) > 5 | int(so_lw_offset_2) > 5)]	as above, but at least one wrist is detected with a confidence score of more than 5 within 2 seconds to the left and right of the word	1,349

Due to the timing information included, we can then download the concordance with links to video snippets that can be loaded into ELAN. A further manual screening stage can be included depending on the research question, e.g. to annotate whether there is gesture on or around the target expression. This process will be briefly described in the following section.

2.5. Rapid Annotator

Screening large numbers of short video snippets is a task that existing multimodal annotation software is often ill-equipped for. The Red Hen Rapid Annotator 2.0 (see Uhrig 2022 for details) is a web-based system that allows researchers to upload video (or audio, image, text) datasets or CQPweb downloads with video snippet links. The experimenter can then define sets of classification questions, e.g. "Is the speaker performing a hand gesture?". Annotators will be presented with the video and the question and can reply with one single keystroke, which will immediately bring up the next question (which can be on the same video or on the next video, which is already preloaded in the background to allow for an instantaneous transition). The results can be downloaded as a spreadsheet and only the relevant videos can then be selected for manual analysis as described in the following sections.

2.6. Manual gesture annotation

We selected video data for annotation and analysis following corpus searches for the verb *will* and hands visible. The video

[1] The attribute *so_speaks* entails *so_detected*, so that the latter can be omitted in this query without change of result.

snippets selected were each 20 seconds long and incorporated not just grammatical markers of future, but also other linguistic markers of time, e.g. time expressions, words with future semantics, or present and past verbs with a future reference. We worked on seven types of 'future' speech markers, but we focus only on the instances of 'will + the infinitive' and the co-occurring hand gesture in this paper. Having performed extensive exploratory manual analysis of 47 video snippets, we translated that analysis into annotations in ELAN for linguistic markers of future, type of those linguistic markers, gestural stroke vs hold, gestural axis, direction, handedness, handshape, or hand orientation[1]. We also automatically annotated for gestural zone: Hands and fingers were annotated for on separate tiers. All linguistic and gestural annotation in ELAN was performed independently from the prosodic annotation.

2.7. Manual prosodic annotation

Prosodic analysis and annotation were carried out using a combination of auditory analysis and close visual inspection of the acoustic waveform and spectrogram, within Praat (Boersma & Weenink 1992-2023) and later incorporated into ELAN[2].

The audio files were annotated in a seven-tier text grid in Praat, consisting of tiers labelled Phrase, Intonational Phrase, intermediate phrase, Prosodic Word (ProsWord), Accent, Syllable, and Comments. The files were segmented into Intonational Phrases (IPs) and Intermediate Intonational Phrases (ips), the identification of which was based on the identification of nuclear pitch accents, an evaluation of tonal sequences, and the identification of boundaries (through cues such as lengthening, segmental strengthening, pauses, the presence of phrase and boundary accents), and meaning. Pauses, both filled and unfilled, were segmented out on these tiers and labelled as 'FP' and 'P' respectively. An orthographic transcription of the contents of each IP was given in the Phrase tier, for ease of reference. The files were further segmented into prosodic words on the ProsWord tier, according to the realization of the utterance in question. Pitch accents, phrase accents, and boundary tones were marked on the Accent tier, a point tier, using IViE (Grabe et al. 1998) conventions. A particular focus was given to nuclear stressed syllables, which were segmented out on the Syllable tier and labelled as 'N'. A final tier was provided for comments, where elements of particular interest were noted, such as mispronunciations, interruptions, speech rate discontinuities, strong focal emphasis, or voice quality effects. The annotation of prosody was performed without access to the linguistic and gestural annotations described in the previous section.

3. Case Study

Having completed our manual and automatic annotation in ELAN for speech (text and prosody) and the gestural features under consideration, we progressed to analysing text-prosody-gesture relations this time using ELAN as a tool for analysis, as it presented us with the ability to see various multimodal features simultaneously, for a given moment or time interval. More specifically it enabled us to engage with the following questions, among others, using our empirical analysis of 47 video snippets depicting future events and scenarios: What is gesture? Is gesture an individual stroke, or a number of strokes perceived as a whole impressionistically, or a sequence of gestural strokes between two positions of rest?

Since our study considers temporal – future – gesture at the level of longer stretches of discourse, the problem of identifying the boundaries of gestures or gestural units becomes acute for both multimodal annotation and analysis.

Temporal gesture belongs to the class of representational gestures as defined by Chu et al. as those that "depict a concrete or abstract concept with the shape or motion of the hands (iconic gestures and metaphoric gestures in McNeill 1992), or point to a referent in the physical or imaginary space (concrete or abstract deictic gestures in McNeill, 1992)" (Chu et al. 2014: 2).

We worked to determine boundaries of gestural units following the Information Packaging Hypothesis, which "states that gesturing helps the speaker organize information in a way suitable for linguistic expression" (Kita 2000: 180) with the process of organising information relying on collaboration between the speaker's analytic and spatio-motoric thinking. We hypothesised that as discourse unfolds this collaboration manifests itself in a dialogue between textual, prosodic, and gestural modalities. We predicted that if we compared boundaries of *impressionistically* perceived gestural boundaries not just with boundaries of linguistic units – words, expressions, clauses, and sentences – but also with the boundaries of prosodic words and phrases, that should help us determine the boundaries of gestural units in a better-informed way.

The results of the corresponding analysis done so far demonstrated that

1. gestural sequences composed of more than one gestural stroke were *impressionistically* perceived as co-occurring with linguistic units composed of more than one word in time conceptualisation;
2. such gestural sequences are almost always composed of strokes[3], made on more than one axis, e.g. on the clause 'what kind of architecture people will need in the future' the *prevailing* axis is sagittal, but we also observe the engagement of the vertical axis on 'will need in the future', when both hands move upwards. See Example 1 (click or scan the QR code).[4]
3. the prevalence of one axis in a gestural sequence – which is complex – serves as a formal indicator that that gestural sequence forms a gestural unit, larger than a stroke but smaller than gestural sequences separated by the position of rest. In our example this axis is sagittal, and we call the respective gestural unit an *overarching temporal gesture*.

[1] All annotations for hand gesture were made by a minimum of two expert coders. All disagreements were discussed and resolved. In especially complex cases a third and sometimes fourth expert coder got involved.

[2] All annotations for prosody were made by two coders. All disagreements were discussed and resolved. In especially complex cases a third expert coder got involved.

[3] among other gestural features, such as gestural holds, changes in hand orientation or shape, etc.

[4] The examples comprise the video snippets and demos of the respective ELAN files with manual and computer-generated annotations.

4. boundaries of such perceived overarching temporal gestures coincide with boundaries of (prosodic) intermediate phrases for the depictions of futures studied here.
5. in cases where it is hard to determine the prevailing axis in what is impressionistically perceived as one unit of overarching temporal gesture the engagements of two axes are separated by a (prosodic) pitch accent. At the same time, these two axes are still conceptually united by one intermediate phrase.
6. the smaller gestural units which form part of the overarching temporal – future – gestures tend to be made along the vertical axis. Those small gestures do not have to fall within the boundaries of the respective linguistic (textual) markers for futures. Rather they co-occur or overlap either with the prosodic word which often incorporates the latter or with intervals which are created between two pitch accents or a boundary tone and an actual boundary of the prosodic phrase as speech unfolds.

   For Example 1, the vertical axis is engaged within the intervals created between two accents which in turn fall within linguistic sequences: ‘an architect[1] should’, ‘should be able to understand what is happening’, ‘predict’, ‘architecture people’, ‘people will need in the future’. In Example 2 (click or scan the QR code), the vertical axis is engaged within the boundaries of the prosodic words ‘futurist’, ‘how long’, ‘will last’, and ‘in the 21st’. The engagements of the vertical axis occur against the backdrop of the overarching temporal gestures made along the sagittal and lateral axes, which prevail.
7. the gestural moves along the vertical axis tend to be rather small in amplitude. They can manifest themselves through the movement of hand, finger(s), or both. For examples, on the prosodic words ‘futurist’ and ‘how long’, we observe very small and brief index finger moves upwards. On ‘in the 21st’, the right hand moves upwards briefly forcing the index finger to move up, too. On ‘will last’ the right hand briefly moves upwards with the fingers pointing up as the hand goes back to the centre along the lateral axis. The engagements of the vertical axis against the background of the lateral or sagittal axes which prevail in examples 1 and 2 constitute further segmentation of the gestural sequences which we observe at the higher level of abstraction – overarching temporal gestures. Further segmentation of the intermediate phrases at the higher level of abstraction for prosody into prosodic words or intervals created by accents appears to be in a dialogue with this further segmentation of the overarching temporal gesture. Smaller units of both gestural and prosodic modalities reinforce each other to support better packaging of information in future conceptualisations. Further segmentation at the prosodic level – into intervals smaller than intermediate phrases – also assists us in determining the boundaries of smaller gestural units.
8. a few examples for which the overarching temporal gesture is made along the vertical axis with the choice of axis were motivated by the semantics of the content verb. For those cases we observed the engagement of the secondary axis too. E.g. for ‘prices will go up’, ‘there will be a hike in it’, there were the engagements of vertical and sagittal axes.

The incorporation of the prosodic analysis also helps to determine boundaries between an outward-directed gesture (ODG) and a body-directed gesture (BDG; on the body-directed gesture see e.g. Wilson 2020). In Example 2 on ‘So being a futurist in a way, how long do you think parametricism will last in the 21st century?’, the speaker’s right hand makes a BDG which co-occurs with ‘So being a futurist in a way, how long do you think’. On ‘parametricism will last in the 21st century’ the right hand goes far rightwards and then returns to the centre before transforming into the BDG on ‘in the 21st century’.

So where exactly do boundaries between the overarching temporal gesture performed along the lateral axis and two BDGs – performed before and after that temporal gesture – lie? Does the first BDG start when the right hand starts moving or does it start when the hand is already moving along the lateral axis? Equally, does the second BDG start when the right hand starts moving back from its far-right position or does it start when the hand is located much closer to the Speaker’s head just milliseconds before the hand touches the Speaker’s chin? Determining the boundaries of the overarching temporal gesture (ODG) through its coinciding with the boundaries of an intermediate phrase helps to determine the boundaries of BDGs.

The process of comparing the gestural boundaries and prosodic boundaries in our data was not mechanical. On the contrary, in making these comparisons we looked at the dialogue between units observed for all three modalities and assessed discourse unfolding from the perspective of information packaging. It was clear to us that the division of the sentence under consideration in at least three overarching gestures – BDG, ODG, and BDG – was motivated by the underlying conceptual blend, core to which is the relation between future and present (for conceptual blending see Fauconnier & Turner 2002; Fauconnier & Turner 2008). From the information packaging perspective, we have a futurist predicting the future while being in the present, parametricism which is already there in the present, but will last in the future, and the mention of the 21st century, which is a temporal space of the present event and the future event at the same time. The sentence with its gestural and prosodic arrangements works to package information which is conceptually complex into gestural and prosodic units whose boundaries coincide. Those units reinforce each other and support the conceptual relation of part-whole as far as the temporal blending for future-present is concerned. We observed the coherence between gesture and prosody in the Speaker’s packaging of information, but it still remains an open question whether one modality leads, and if so, which.

4. Conclusion and New Directions

The limitations of space prevented us from presenting all the interactions between the manual and computational work that we have engaged in and which was necessary to be able to efficiently carry out research on the interaction of all three modalities.

We have shown that the computer-assisted retrieval and subsequent analysis of prosody and gesture allowed us to establish the coherence between gesture and prosody in respect to the boundaries of gestural and prosodic phrases. More specifically, we observed the co-occurrence of the boundaries of overarching temporal gestures and intermediate prosodic phrases. Looking into the boundaries of lower-level prosodic features helped us to determine the engagement of a secondary

[1] ‘an architect’ has a future reference in the context as s/he is defined as a futurist as discourse unfolds.

axis within the overarching temporal gesture. It is only the combination of qualitative and automatic methods that enabled us to reveal this pattern. The pattern will need to be examined using data from other speakers before we can generalise further and assess our findings from a theoretical perspective.

As a next step, we plan to cluster vector representations of hand movements and facial gestures extracted from pretrained deep neural networks combined with audio features, with the aim to find natural structure behind gestures. We hypothesize that the resulting robust automatic methods of multimodal data analysis together with explainable AI methods will help us determine what the machine sees as gesture. The combination of our human and machine approaches should ultimately help us determine what temporal gesture is.

5. Acknowledgements

The authors would like to thank Georgy Krasovitsky & Anna Saputsova for help with annotations in earlier stages of the research project.

The research presented in this paper was made possible by generous funding provided by the Deutsche Forschungsgemeinschaft (project number 468466485) and the Arts and Humanities Research Council (grant reference AH/W010720/1). The authors gratefully acknowledge the scientific support and HPC resources provided by the Erlangen National High Performance Computing Center (NHR@FAU) of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) under the NHR project b105dc. NHR funding is provided by federal and Bavarian state authorities. NHR@FAU hardware is partially funded by the German Research Foundation (DFG) – 440719683.

6. References

Alam, T., Khan, A., & Alam, F. 2020. Punctuation Restoration using Transformer Models for High-and Low-Resource Languages. *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, ACL, 132–142.

Boersma, P. & Weenink, D. 1992–2022. *Praat: doing phonetics by computer [Computer program]*. Version 6.2.09, retrieved 2 March 2023 from https://www.praat.org.

Cao, Z., Hildago, G., Simon, T., Wei, S. & Sheikh, Y. 2018. *OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields.* arXiv preprint, arXiv:1812.08008.

Chu, M., Meyer, A., Foulkes, L. & Kita, S. 2014. Individual differences in frequency and saliency of speech-accompanying gestures: The role of cognitive abilities and empathy. *Journal of Experimental Psychology: General* 143.2, 694-709.

Cooperrider, K., Núñez, R. & Sweetser, E. 2014. The conceptualization of time in gesture. In: *Body-language-communication* 2, 1781-1788.

Grabe, E., Nolan, F., and Farrar, K. 1998. IViE - A comparative transcription system for intonational variation in English. *Proceedings of ICSLP* 98, Sydney, Australia.

Fauconnier, G. & Turner, M. 2002. *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities*. New York: Basic Books.

Fauconnier, G. & Turner, M. 2008. Rethinking metaphor. In: Ray Gibbs (Ed.) *Cambridge Handbook of Metaphor and Thought*. New York: Cambridge University Press, 53–66.

Hardie, A. 2012. CQPweb: Combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics* 17(3), 380–409.

Kita, S, 2000. How representational gestures help speaking. In McNeill, D. (Ed.) *Language and Gesture*. Cambridge: Cambridge University Press, 162–185.

McNeill, D. 1992. *Hand and mind: What gestures reveal about thought*. Chicago & London: University of Chicago Press.

Pouw, W. & Trujillo, J. P. 2021. *Selecting, smoothing, and deriving measures from motion tracking, and merging with acoustics and annotations*. Retrieved 2 March 2023 from: https://wimpouw.github.io/EnvisionBootcamp2021/MergingAcousticsMT.html

Proisl, T. & Uhrig, P. 2016. SoMaJo: State-of-the-art tokenization for German web and social media texts. *Proceedings of the 10th Web as Corpus Workshop*, Berlin: ACL, 57–62.

Radford, A., Kim, J., Xu, T., Brockman, G., McLeavey, C. & Sutskever, I. 2022. *Robust Speech Recognition via Large-Scale Weak Supervision*. arXiv preprint, arXiv:2212.04356.

Straka, M. & Straková, J. 2017. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, Vancouver, Canada, August 2017, 88–99.

Tao, R., Pan, Z., Das, R. K., Qian, X, Shou, M.Z. & Li, H. 2021. Is Someone Speaking? Exploring Long-term Temporal Features for Audio-visual Active Speaker Detection. *Proceedings of the 29th ACM International Conference on Multimedia*, 3927–3935.

Uhrig, P. 2022. *Large-Scale Multimodal Corpus Linguistics – The Big Data Turn*. Habilitation Thesis, FAU Erlangen-Nürnberg.

Wilson, A. 2020. It's Time to Do News Again. *Zeitschrift für Anglistik und Amerikanistik* 68, 379–409.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A. & Sloetjes, H. 2006. ELAN: A professional framework for multimodality research. *Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation*. Genoa, Italy: ELRA, 1556–1559.

Authors' contributions:

Peter Uhrig: development of the overall research design, leading on the development of computational tools, and writing the paper

Elinor Payne: leading on the analysis of prosody; writing the paper's section on prosody annotation

Irina Pavlova: annotation of data for speech and hand gesture; annotation for speech-gesture interaction

Ilya Burenko: development of computer vision pipeline and drafting of the corresponding section in the paper

Nathan Dykes: contribution to the development of the NLP pipeline, corpus management and drafting of the corresponding section in the paper

Mary Baltazani: advising on the annotation scheme for prosody and exploratory analysis for prosody

Evie Burrows: annotation for prosody, writing the paper's section on prosody annotation

Scott Hale: contribution to and maintenance of the computational infrastructure supporting the analysis

Philip Torr: advising on computer vision-related work

Anna Wilson: development of the overall research design, leading on the annotation and qualitative (manual analysis), performing annotation of data for speech and gesture, annotation for and analysis of gesture-prosody interaction, and writing the paper