

Automatic Annotation of Co-speech Gesture for Multimodal Communication Research – A Framework and a Case Study

Ilya Burenko^{1,2}, Peter Uhrig^{3,1}, Anna Wilson⁴

¹ScaDS.AI Dresden/Leipzig, Germany*

²Technische Universität Dresden, Dresden, Germany

³Friedrich-Alexander Universität Erlangen-Nürnberg, Erlangen, Germany

⁴Oxford School of Global and Area Studies, University of Oxford, United Kingdom

Correspondence: ilia.burenko@tu-dresden.de

Abstract

Large-scale multimodal communication research increasingly depends on computational methods capable of extracting reliable gesture annotations from video data. Yet it remains unclear to what extent automatically generated annotations can support the linguistic investigation of naturally occurring co-speech gesture and where expert interpretation remains indispensable. This paper presents a framework for the automatic analysis and annotation of co-speech gesture in media corpora, developed to support multimodal communication research at scale. We build our framework on top of the existing computer vision frameworks for keypoint extraction and active speaker detection, with post-processing steps carefully designed to mitigate noisy output. Building on the processed keypoint trajectories, we implement a configurable rule-based annotation system developed through close collaboration between computer vision researchers and linguists. We demonstrate the usefulness of the framework through a case study investigating vertical hand movement accompanying English expressions of future and past, examining which aspects of this gestural pattern can be captured through automatic annotation and which require expert multimodal interpretation. We find that at corpus scale, greater mean amplitude and higher mean peak velocity converge in *will*-future, suggesting greater kinematic prominence in the multimodal expression of futurity. More broadly, the case study demonstrates how computational pipelines can support, rather than replace, human analysis by enabling multimodal hypotheses to be investigated on datasets that would be impractical to analyse manually.

1 Introduction

The combination of increasingly large multimodal corpora and advances in computational methods is

transforming multimodal communication research. Researchers can now investigate patterns of speech and gesture across thousands of naturally occurring interactions rather than a few dozen manually annotated examples. At the same time, this opportunity presents a major methodological challenge: producing reliable multimodal annotations at such a scale remains prohibitively time-consuming when performed manually, especially when the annotation process requires trained experts.

In linguistics, ELAN (Wittenburg et al., 2006) serves as the standard environment for producing and editing manual annotations of audiovisual recordings. It enables researchers to create multiple annotation tiers, which are stored alongside the media files. ELAN annotations can also be manipulated programmatically using the Python package *pympl-ing* (Lubbers and Torreira, 2013-2026). Larger collections of annotated data can be organised into corpora and queried using corpus managers such as CQPweb (Hardie, 2012).

Although the machine learning methods have achieved substantial improvements in computational performance, comparatively less attention has been paid to how they support multimodal communication research itself. Most existing work evaluates computational performance, whereas comparatively few studies investigate what kinds of linguistic hypotheses automatic annotation can reliably support, where expert multimodal interpretation remains indispensable, and what new research questions become accessible once annotation can be performed computationally at corpus scale.

To demonstrate the usefulness of the framework, we apply it to the NewsScape 2016 corpus (Uhrig, 2022), which contains more than 30,000 hours of American TV news (in a broad sense) with subtitles comprising over 230 million words that have been time-aligned with Gentle (Ochshorn and Hawkins, 2015). Processing a dataset of this size presents substantial computational challenges. Temporal

*Center for Scalable Data Analytics and Artificial Intelligence Dresden/Leipzig

expressions cannot simply be retrieved from raw video but require several computationally demanding processing stages, including scene detection, pose estimation, active speaker detection and keypoint extraction.

Rather than evaluating the framework solely as a technical system, we demonstrate its usefulness through a genuine multimodal communication research question: whether previously identified patterns of vertical gesture accompanying English temporal conceptualisation generalise to a substantially larger automatically annotated corpus. The case study simultaneously investigates what aspects of this phenomenon current automatic annotation can recover reliably, where expert multimodal interpretation remains essential, and what additional linguistic questions become accessible once gesture can be analysed computationally at corpus scale.

The paper therefore makes three contributions. First, it presents a scalable, modular framework for the automatic annotation of co-speech gesture in multimodal corpora. Secondly, through a multimodal case study, it demonstrates how computational annotation can generalise findings previously established through manual multimodal analysis while revealing both the capabilities and the current limitations of automatic gesture annotation. Thirdly, it shows how corpus-scale computational analysis enables new forms of multimodal communication research by quantifying kinematic properties of gesture that would be impractical to investigate manually.

The remainder of the paper is organised as follows. Section 2 reviews related work. Sections 3 and 4 describe active-speaker keypoint extraction, post-processing and the automatic annotation framework. Section 5 validates the automatic annotation against expert human annotation, and Section 6 presents the case study of co-speech gesture accompanying English expressions of future and past. Section 7 discusses the methodological and linguistic implications, followed by limitations of the automatic annotation and analysis in Section 8.

2 Related work

Scene Detection. Several stages of our pipeline require partitioning a video into scenes – segments of footage in which the camera position and filmed content remain relatively stable –, including active speaker detection (Section 3) and keypoint normal-

isation (Section 4). We use PySceneDetect¹ for this purpose, a content-aware detection tool that identifies a scene boundary whenever the rate of change in frame colour and saturation exceeds a fixed threshold. PySceneDetect exposes a range of configurable parameters, but we found the default settings sufficient for our dataset.

Pose Estimation. To detect people in video frames and estimate two-dimensional coordinates of their joints, we need to use a pose estimation method. There is a wide range of readily available pose estimation models, which allows to choose the best trade-off between speed and quality. Light-weight models such as MediaPipe (Lugaresi et al., 2019) can process video with high frame-per-second ratio, sacrificing some quality. In contrast, recent large models such as Sapiens (Khirrodkar et al., 2024) yield better performance, but it comes at an additional computational cost. Among others we compared AlphaPose (Fang et al., 2022), HRNet (Cheng et al., 2020), or OpenPifPaf (Kreiss et al., 2021), and OpenPose (Cao et al., 2019). We chose OpenPose as the optimal tool to process the full 30,000-hour dataset as it can process multiple people on screen, is very robust, and proved to be the best trade-off between speed and quality. We refer the reader to Gao et al. (2025) for a recent review of pose estimation methods.

Active Speaker Detection. To extract OpenPose coordinates for the active speaker, we first automatically identify which individual is speaking at any given time. The AVA Challenge 2022² evaluated a number of approaches for active speaker detection. These employ diverse architectural strategies: some approaches, such as that proposed by (Köprüklü et al., 2021), utilize 3D convolutional networks combined with bidirectional Gated Recurrent Units (Cho et al., 2014) to capture temporal dependencies in the data, while alternative methods leverage graph networks to model spatio-temporal relationships between detected faces across sequential frames (Min et al., 2022). For a more detailed discussion, we refer the reader to the survey by Nur Aisyah Mohd Robi et al. (2024).

For our pipeline, we adopt TalkNet-ASD, which employs a transformer architecture with cross-attention mechanisms (Tao et al., 2021) to achieve robust active speaker detection; the authors’ pub-

¹<https://github.com/Breakthrough/PySceneDetect>

²<https://research.google.com/ava/challenge.html>

licly available implementation³ is straightforward to run, test, and adapt. TalkNet-ASD partitions the input video into scenes and detects a face bounding box for each visible individual in every frame. Bounding boxes with substantial overlap across consecutive frames are merged into what the authors term a *track*. If the pipeline fails to detect a face for a given frame within a track, it linearly interpolates the bounding box coordinates for up to 10 frames by default.

Annotation Pipelines. To the best of our knowledge, there is no comparable framework for the automatic annotation of co-speech gestures. Envisionbox⁴ collects community code snippets and recipes for the application of various computer vision tools and how to exploit their results, but no fully-integrated automatic pipeline is provided.

3 Adapting TalkNet-ASD to Dataset Scale

We adapt TalkNet-ASD (Tao et al., 2021) to accommodate the scale and heterogeneity of large multimodal datasets. Two properties of such datasets are not well supported by the original implementation. First, some videos contain exceptionally long scenes, sometimes extending to several hours; press conference recordings filmed from a fixed camera angle, for instance, produce very long tracking sequences. Second, video collections often span a broad range of resolutions, from low-resolution recordings dating back to the early 2000s to recent 4K content. The original implementation loads all video frames into memory, which is not feasible for high-resolution video.⁵

To address these limitations, we implement three optimisations. High-resolution videos are resized before being loaded into memory, reducing computational overhead. Long scenes and tracks are split into smaller sub-scenes and sub-tracks, which can be processed more efficiently. Finally, unlike the original implementation, which processes a single image per inference step, our version uses a batch size of 32, which accelerates inference considerably. These changes do not alter the underlying algorithm; they enable its use with video of any resolution and scenes of any length.

³<https://github.com/TaoRuijie/TalkNet-ASD>

⁴<https://envisionbox.org/index.html>

⁵Note, however, that the pre-trained network only accepts video at 25 fps. TalkNet-ASD converts any input video to 25 fps before applying the network.

4 ASD Keypoints Extraction

For body pose annotation, we use OpenPose (Cao et al., 2019), from which we extract keypoints in the BODY 25 scheme together with hand and face keypoints. The pipeline is not tied to OpenPose specifically: keypoints produced by any other pose estimation model can be used in its place, provided they are mapped onto the BODY 25 scheme, which is generally straightforward.

Since OpenPose estimates keypoints for every person visible in a frame, we need a way to pick out the active speaker among them. To identify which of these predictions corresponds to the active speaker, we use the bounding box predicted by TalkNet-ASD: for each frame, we select the individual whose head position falls within this bounding box.

The resulting keypoint time series are noisy in several ways. Keypoints are sometimes missing entirely, for example when motion blur or occlusion prevents OpenPose from detecting a body part. The coordinates also jitter slightly from frame to frame, even when the corresponding body part is not actually moving. Finally, because OpenPose predicts coordinates in raw pixel values, the same body part can appear at very different scales depending on the camera angle and the speaker’s distance from the camera. Left uncorrected, these issues make the data difficult to work with, both for automatic analysis and for manual inspection. We therefore apply a three-step preprocessing pipeline, separately for each track (Section 2): interpolation and extrapolation, smoothing, and normalisation.

Missing values are handled through interpolation and extrapolation. Gaps between two valid measurements are filled by linear interpolation, while missing values at the start or end of a track are filled by extrapolating the nearest valid value. Further details are given in the Appendix C.

Jitter is reduced through smoothing. Each coordinate is replaced with a weighted average over a sliding window of 11 frames, centred on the current frame, using Epanechnikov kernel weights. This window size offered the best trade-off in our data between removing jitter and preserving genuine small movements.

To correct for scale, we normalise the coordinates into a shared coordinate space that is independent of the speaker’s size and position in the frame. Body keypoints are centred on the nose and scaled by the track-averaged distance between the nose

and neck; face keypoints are centred on the tip of the nose and scaled by the track-averaged distance between the nasal bridge and the nose tip. Hand keypoints are left unnormalised, since OpenPose’s hand predictions are comparatively unreliable to begin with.

Figure 1 shows the pipeline’s effect: the top panel shows raw pixel coordinates, and the bottom panel shows the same time series after interpolation, extrapolation, smoothing, and normalisation. Implementation details, including the keypoint indices used and the reasoning behind our normalisation choices, are provided in Appendix C.

5 Annotation Validation

To validate the pipeline for our case study, which focuses on the vertical axis, we compared the automatic annotation with human annotations for the same data (24 episodes of SophieCo. Visionaries show; see Wilson et al. (2024) for details of the dataset). The pipeline detected upward and downward movements independently for each wrist. Human annotations specified the direction of movement (UP/DOWN) and the hand involved (Right, Left, or Both). Table 1 presents the results of the comparison between manual and automatic annotations. No overlap denotes human annotations for which no overlapping automatic annotation was found. Where the annotations overlap, Strict denotes agreement in movement direction for all annotated hands, Relaxed agreement for one hand, and Wrong no agreement in direction. For single-hand human annotations, a machine annotation is classified as Strict only when it identifies the same direction for the same hand and detects no movement for the other wrist.

The validation videos comprise Studio and Home recordings, which differ in lighting, camera angle, and video quality. These recording conditions affect the pipeline’s ability to detect movements visible to human annotators, as reflected in the higher No-overlap rate for Home recordings. Once No-overlap cases are excluded, however, the proportions of Strict, Relaxed, and Wrong classifications are comparable across the two conditions.

6 Case Study

6.1 Previous findings and research questions

Wilson (under review) investigated how English speakers conceptualise time through co-speech ges-

ture using manual multimodal annotation, statistical analysis, and conceptual blending analysis. The study demonstrated that English temporal conceptualisation is realised through a multidirectional gestural system rather than being restricted to the canonical forward-backward or rightward-leftward directions. Gestural direction was shown to be motivated by meaning construction in discourse, gesture multifunctionality and “gestural syntax” – a concept proposed by Wilson –, with sagittal, lateral and vertical directions jointly contributing to temporal meaning. The quantitative analysis of 151 manually annotated future and past speech–gesture instances identified the vertical axis as a discriminative parameter within the grammatical constructions examined: *will*-future constructions were predominantly accompanied by upward movement or positioning, whereas present perfect and simple past constructions were predominantly accompanied by downward movement or positioning. Qualitative analysis further demonstrated that the communicative function of vertical movement can only be interpreted in relation to the surrounding gesture sequence and discourse context rather than as an isolated movement.

The present case study investigates whether these findings generalise beyond the manually analysed 151 instances and how automatic annotation can support or extend such research. Specifically, we ask three related questions.

1. Does the systematic engagement of the vertical axis remain observable when vertical hand movement is analysed automatically across a substantially larger dataset?
2. What does this case study reveal about the validity and usefulness of automatic gesture annotation for multimodal communication research?
3. What new linguistic insights become possible once gesture can be analysed computationally at corpus scale?

To address these questions, we ultimately engaged with more than 80,000 speech units representing *will*-future, present perfect and past perfect to explore hand movement accompanying these temporal constructions in English.

6.2 Quantitative analysis

The NewsScape 2016 corpus was queried for combinations of *will* followed by one of the following

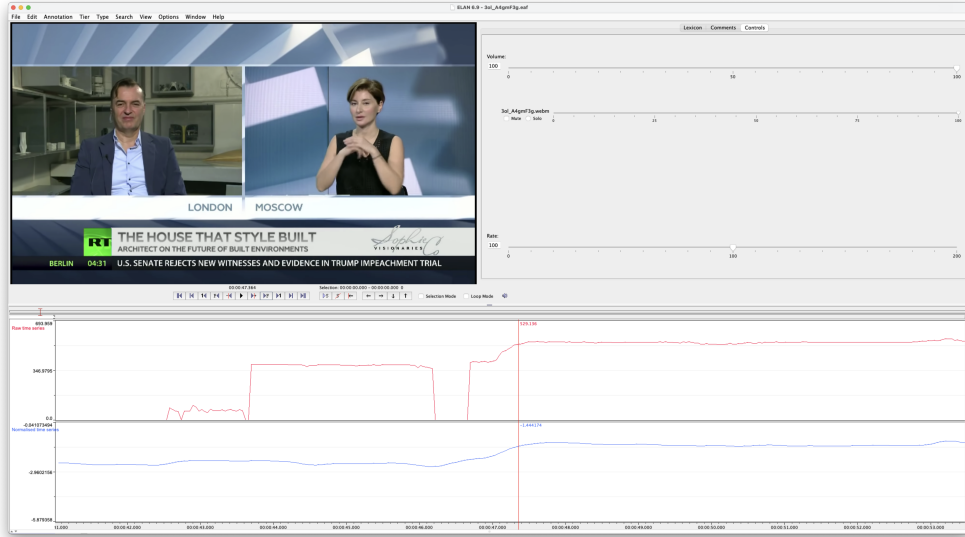


Figure 1: Time series of the vertical (y-)coordinate for the right wrist of the active speaker. The red time series (top panel) shows the raw OpenPose coordinates; zero values correspond to failures in detecting coordinates. The blue time series shows the post-processed (normalised and interpolated) version of the same signal, with less jitter and no missing values. The change in scale for the red line at time code 00:00:43.6 is due to a scene change. The gap at 00:00:46.2–00:00:46.6 is where OpenPose fails to predict the wrist’s coordinates, visible as a drop in the raw signal but absent in the interpolated one.

Location	N	No overlap	Strict	Relaxed	Wrong
Home	329	152 (46.2%)	62 (18.8%)	66 (20.1%)	49 (14.9%)
Studio	99	12 (12.1%)	34 (34.3%)	33 (33.3%)	20 (20.2%)
Total	428	164 (38.3%)	96 (22.4%)	99 (23.1%)	69 (16.1%)

Table 1: Results of the annotation validation

verbs, which were inspired by [Wilson \(under review\)](#): *decrease*, *go down*, *increase*, *go up* (spatial: upwards and downwards), *stop*, *terminate*, *delete* (non-spatial: stopping), *read*, *swim* (non-spatial: neutral), and for sequences of *have*, *has* or *had* and the past participle of the same verbs. This yielded 7,926 hits, 839 of which were discarded because at least one of the words was not successfully aligned with the audio (see [Uhrig \(2022\)](#) for details). A further 84 instances with implausible lengths (less than 4 frames – approx. 133 ms or more than 90 frames – approx. 3 s) were removed. A further 2,914 instances were removed because no active speaker was detected. 2,408 hits were then removed because no keypoints were detected or no wrist had an average confidence score of at least 0.3. Out of the remaining 1,681, 579 showed no wrist movement above a small threshold used to filter resting hands and longer holds. Overall, more than 85% of query results were excluded, consistent with previous manual annotation of media data.

For the remaining 1,102 hits, wrist movements were counted separately for the auxiliary, content verb, and both; movements extending beyond the two-word span were excluded to minimise the influence of surrounding words.

The quantitative analysis demonstrates that the vertical axis is systematically engaged in all three grammatical constructions. Across the constructions, the auxiliary favours UP movement and the content verb DOWN movement. For movements spanning both positions, DOWN movements outnumber UP movements in *will*-future, UP movements slightly outnumber DOWN movements in present perfect, while past perfect is evenly divided (Table 5). These findings confirm that the engagement of the vertical axis identified by [Wilson \(under review\)](#) generalises to a substantially larger automatically annotated corpus, although the automatic UP/DOWN distributions show considerable overlap.

The overlap in the automatic UP/DOWN distributions could reflect limitations of the automatic

annotation pipeline, the complexity of naturally occurring multimodal communication, or the operationalisation of movement within predefined temporal windows.

To investigate these possibilities, we conducted a qualitative evaluation of 100 randomly selected clips from the automatically annotated dataset. Each clip was independently annotated manually in the Red Hen Rapid Annotator (Uhrig, 2022) by two expert human annotators and compared with the machine annotation. For every example we recorded observations concerning the quality of the automatic annotation together with its multimodal interpretation.

For a broader quantitative analysis, we queried the 102 verbs listed in Appendix A with *will*, *have*, *has* and *had*, and in the third-person singular present and past tense. The verbs were selected for their lack of direct temporal or spatial semantics. Random samples of 20,000 hits were extracted for *will*-future, present tense, present perfect and past tense, together with all 7,021 past-perfect hits. For each wrist movement, vertical amplitude was calculated as the difference between the y-coordinates of the last and first keypoints. Positive and negative values represent UP and DOWN movement, respectively (= signed amplitude). We also calculated absolute amplitude, i.e. the absolute value of the vertical distance. Peak velocity was calculated for combined lateral (*x*) and vertical (*y*) axes, and vertical axis, because the latter on its own does not indicate how fast the hand moves. Velocities for both types of measurements were calculated as the maximum change in coordinates between two consecutive wrist keypoints, multiplied by the frame rate.

The expanded quantitative analysis reveals differences between the three constructions. *Will*-future has the highest mean peak velocities for both UP and DOWN movements at the auxiliary, content-verb and combined positions (Table 3), and the largest absolute vertical amplitude at all three positions (Table 4). For absolute amplitude, comparisons between *will*-future and perfect constructions after movement-length normalisation were statistically significant in all cases except *will*-future versus past perfect at the auxiliary and content verb (Table 2). Absolute vertical amplitude increases from auxiliary to content verb in all three constructions, although this may be partly conditioned by our focus on movements within predefined temporal windows, and the longer duration of content

verbs relative to auxiliaries (mean length auxiliaries: 169 ± 80 ms; content verbs: 196 ± 94 ms). UP peak velocities remain essentially stable in *will*-future, but decline in present perfect and past perfect (Table 3). This pattern is of interest given that DOWN peak velocities increase from auxiliary to content verb in all three constructions. Our analysis of mean signed amplitude confirmed that across the constructions the auxiliary favours UP movement and the content verb DOWN movement. The differences are statistically significant in all three constructions (Table 6, the raw data in Appendix B).

6.3 What does the case study reveal?

Research Question 1: Does the vertical-axis hypothesis generalise?

The quantitative analysis confirms that the systematic engagement of the vertical axis identified by Wilson (under review) generalises to a substantially larger automatically annotated corpus.

This study's qualitative analysis further supports Wilson's (under review) interpretation of the communicative function of the vertical axis. Human annotators evaluated each temporal grammatical construction as a single multimodal unit rather than treating the auxiliary and verb independently. Consequently, an upward movement or upward hand position associated with the *will* + infinitive construction was counted as positive evidence irrespective of whether it occurred on the auxiliary itself or began slightly before or ended shortly after the construction while remaining temporally associated with it. The latter cases primarily reflected the natural timing of gesture production, including instances where the articulatory onset of the gesture preceded the acoustic onset represented by subtitle alignment. Applying these criteria, 32 of 40 *will*-future constructions (80%) displayed the predicted upward pattern.

The same principles were applied to *has/have/had* + past participle constructions. Downward movement or downward positioning was counted irrespective of whether it occurred on the auxiliary itself (e.g. hands positioned down on *has*), on the participle while clearly belonging to the grammatical construction (e.g. left hand down on *gone down*; left hand up and down on *increased*), or began slightly before or ended shortly after the construction owing to the natural temporal coordination between gesture and speech. Applying these criteria, 30 of 39 constructions

Tenses	Position	Welch's t	p_{Bonf}
<i>Will</i> -future vs Present perfect	Auxiliary	3.2	0.009
	Content verb	4.4	5.54×10^{-5}
	Both	5.3	6.99×10^{-7}
<i>Will</i> -future vs Past perfect	Auxiliary	1.6	0.709
	Content verb	2.7	0.042
	Both	4.7	2.31×10^{-5}

Table 2: Welch's t -statistics on mean length-normalized vertical absolute amplitudes for *will*-future versus the two perfect constructions by verb segment; p -values are Bonferroni-corrected for six comparisons.

Position	<i>Will</i> -future		Present perfect		Past perfect	
	x, y	y only	x, y	y only	x, y	y only
Auxiliary – UP	1.59 ± 1.33	1.24 ± 1.02	1.35 ± 1.12	0.97 ± 0.86	1.48 ± 1.08	1.02 ± 0.77
Auxiliary – DOWN	1.33 ± 1.06	0.94 ± 0.84	1.24 ± 1.20	0.85 ± 0.72	1.13 ± 0.92	0.82 ± 0.68
Content verb – UP	1.58 ± 1.27	1.20 ± 1.12	1.25 ± 1.09	0.91 ± 0.76	1.23 ± 1.04	0.83 ± 0.68
Content verb – DOWN	1.45 ± 1.22	1.10 ± 1.02	1.36 ± 1.23	1.02 ± 1.03	1.38 ± 1.01	0.98 ± 0.72
Both – UP	1.97 ± 1.61	1.54 ± 1.35	1.87 ± 1.84	1.48 ± 2.11	1.96 ± 2.78	1.46 ± 2.55
Both – DOWN	2.07 ± 1.52	1.63 ± 1.37	1.70 ± 1.33	1.30 ± 1.44	1.76 ± 1.42	1.27 ± 1.02

Table 3: Mean peak velocity $\pm$ one standard deviation by position and construction, computed from the x and y (2D) coordinates and from the y coordinates only

(77%) displayed the predicted downward pattern.

The remaining (negative) examples demonstrate why temporal meaning cannot always be recovered from movement counts alone. Individual movements simultaneously fulfil multiple communicative functions, contributing not only to temporal reference but also to event depiction, discourse organisation, stance and interaction. Moreover, temporal interpretation frequently depends on the unfolding gesture sequence rather than on an isolated movement. The qualitative analysis therefore reinforces broader theoretical claim from Wilson (under review) that gestural direction is determined by multimodal meaning construction, gesture multifunctionality (see e.g. Kok et al., 2016) and gestural syntax rather than by fixed one-to-one correspondences between individual movements and linguistic forms (on gesture-speech timing see e.g. ter Bekke et al., 2024; Fritz et al., 2021; and cited literature).

Research Question 2: What does the case study reveal about automatic gesture annotation?

The comparison between machine and human annotation demonstrates the sufficient reliability of the automatic pipeline for large-scale analysis. Of the 100 annotated clips, 15 had to be discarded and excluded from the evaluation either because of errors in the subtitles or because of repetitions in the dataset. Of the remaining 85 instances, 35

(41.2%) showed complete agreement between machine and human annotation, while a further 32 (37.6%) showed partial agreement. Thus, 78.8% of the evaluated examples contained information produced by the automatic pipeline that was fully or partially consistent with expert manual annotation.

Many disagreements arose from factors external to the movement-detection algorithm itself, including partial hand visibility, occlusion by objects or television graphics, subtitle and forced-alignment errors, and camera movement. These limitations are characteristic of naturalistic broadcast data rather than of the computational method itself.

A further source of divergence between automatic and human annotation is that the former tracks wrist trajectories, whereas human annotators interpret the hand as a whole. The wrist and the fingers or palm may move or orient in different directions, producing different directional classifications of the same gesture.

Other discrepancies reflect fundamental properties of human communication. The automatic pipeline detects movement trajectories, whereas human annotators interpret communicative function. Every upward movement necessarily gives rise to a subsequent downward movement, and vice versa. Consequently, the movement detected automatically is not always the movement interpreted by human annotators as temporally foregrounded.

Position	<i>Will</i> -future	Present perfect	Past perfect
Auxiliary	.152 $\pm$.183	.131 $\pm$.158	.137 $\pm$.141
Content verb	.190 $\pm$.223	.160 $\pm$.198	.163 $\pm$.217
Both	.361 $\pm$.355	.315 $\pm$.330	.290 $\pm$.306

Table 4: Mean absolute amplitude $\pm$ one standard deviation by position

Position/direction	<i>Will</i> -future	Present perfect	Past perfect
Auxiliary – UP / Total	482 / 826 (58.4%)	403 / 645 (62.5%)	123 / 183 (67.2%)
Content verb – DOWN / Total	884 / 1,522 (58.1%)	783 / 1,385 (56.5%)	228 / 393 (58.0%)
Both – UP / Total	909 / 2023 (44.5%)	814 / 1555 (52.3%)	230 / 460 (50.0%)
Both – DOWN / Total	1114 / 2023 (55.5%)	741 / 1555 (47.7%)	230 / 460 (50.0%)

Table 5: The fraction of UP or DOWN movements for the type of the verb by tense

Similarly, gestures frequently begin before or continue after the associated auxiliary, so the relevant movement may fall outside the grammatical construction selected for automatic analysis.

We experimented with temporal buffers around the target construction to capture movements beginning before or ending after the auxiliary. Although this approach successfully retrieved additional temporally relevant movements, it also admitted numerous unrelated gestures belonging to neighbouring discourse segments, thereby reducing the specificity of the automatic annotation. Increasing the temporal window therefore does not, by itself, solve the problem, but can be useful to improve recall (i.e. how many gestures are found) at the cost of precision (i.e. the rate of relevant gestures among those retrieved by the machine).

The qualitative evaluation additionally revealed several examples in which the automatic pipeline detected genuine movements that had initially been overlooked by human annotators. Subsequent detailed inspection confirmed that these movements were present in the video. Automatic annotation therefore does not simply reproduce manual analysis but can complement it by identifying subtle movements that may be missed because of attention limitations or annotation fatigue.

Research Question 3: New linguistic insights through automatic annotation at corpus scale
The case study demonstrates that the engagement of the vertical axis identified through manual multimodal analysis generalises to a substantially larger automatically annotated corpus. *Will*-future has the largest absolute vertical amplitude across the auxiliary, content-verb and combined positions, and also the highest mean peak velocity for both upward and downward movements. This convergence

is consistent with research linking gestural prominence to movement magnitude and other kinematic properties (Ferré (2018); Pouw and Dixon (2019)). In our data, the convergence of amplitude and peak velocity therefore suggests that the greater amplitude associated with *will*-future reflects more than simple spatial displacement and may index greater kinematic prominence with which the future is introduced into the multimodal communication (cf. Horst et al. (2014) on movement quality and communicative effort).

The construction-internal peak-velocity profiles show a further descriptive difference between *will*-future and the perfect constructions. DOWN peak velocity increases from auxiliary to content verb in all three constructions, while UP peak velocity remains essentially stable in *will*-future but declines in present perfect and past perfect (see Table 3). This pattern raises the possibility that *will*-future maintains a high-velocity upward component across the construction; however, the substantial variation within the data means that establishing whether this reflects genuine construction-internal stability requires more detailed quantitative and qualitative analysis.

7 Discussion and Conclusions

As we have shown, the case study demonstrates that automatic annotation is most valuable not as a replacement for expert multimodal analysis but as a complementary research instrument. The computational pipeline makes it possible to analyse thousands rather than hundreds of gesture instances, substantially reduces annotation effort, provides objective and reproducible movement measurements, and can identify genuine movements overlooked during manual annotation. Human analy-

Case	N (Aux)	N (Content)	Mean (Aux)	Mean (Content)	Welch t	Welch p
Will-future	826	1522	0.03 ± 0.24	-0.032 ± 0.29	5.7	1×10^{-8}
Present perfect	645	1385	0.035 ± 0.2	-0.038 ± 0.25	7.0	4×10^{-12}
Past perfect	526	1127	0.058 ± 0.18	-0.068 ± 0.27	11.2	6×10^{-27}

Table 6: Welch’s t -test results by case

sis remains essential for interpreting communicative meaning within the unfolding multimodal discourse, whereas automatic annotation excels at detecting, measuring and retrieving gesture at scale. Together, the two approaches provide a scalable methodology for multimodal communication research that neither could achieve independently. If a sufficiently large dataset is available, the noise inherent in our system becomes manageable so that patterns can be discerned from fully automatically annotated data without human intervention in the annotation process, although we recommend always manually verifying a sample to confirm the patterns identified by the automatic analysis. Based on our analysis of a relatively large dataset, we can confirm the hypothesis from (Wilson, under review) that the vertical axis plays a role in the expression of time, while the combined automatic and qualitative analyses support the finding that *will*-future favours upward movement.

Linguistic contribution. Our case study demonstrates how automatic annotation can generate new linguistic questions that would not be practical to investigate through manual annotation alone. The ability to quantify properties such as movement amplitude and velocity across large samples makes it possible to investigate how gesture kinematics and multimodal prominence relate to the contrasting conceptualisation of future and past in speech–gesture communication. This extends previous work on multimodal prominence and the role of gesture in conceptualisation and information structuring into the domain of temporal conceptualisation at corpus scale (on multimodal prominence see Krahmer and Swerts, 2007; Swerts and Krahmer, 2008; Kita, 2000; Kita et al., 2017; Wilson, 2020 and the cited literature).

Methodological contribution. The case study demonstrates that computational pipelines and qualitative multimodal analysis answer complementary research questions. Automatic annotation captures movement trajectories at corpus scale reliably enough, whereas qualitative analysis establishes their communicative significance. Rather than re-

placing expert interpretation, computational annotation enables new forms of multimodal inquiry by combining large-scale quantitative analysis with detailed qualitative investigation.

8 Limitations

Automatic annotation and human multimodal analysis differ in what they capture. The automatic pipeline tracks the trajectories of individual wrists within predefined temporal windows, whereas human annotators perceive the hand as a whole, including the palm and fingers, and can interpret the movements of both hands together. Human annotators can also identify which movement is perceptually foregrounded and interpret it in relation to preceding and subsequent gesture and speech. The automatic analysis cannot make these contextual distinctions and restricts analysis to the selected temporal window, even though relevant gestures may begin before or continue beyond it.

The pipeline is also subject to technical limitations arising from keypoint extraction and normalisation. Partial hand visibility, occlusion and recording quality can impair keypoint detection, while camera movement creates additional problems for normalisation. In dynamic scenes, substantial camera movement can bias the reference coordinates used for normalisation, leading to shifted body coordinates. Frame-by-frame coordinate subtraction could reduce this problem, but may introduce apparent wrist movement when the wrist remains stationary while the speaker’s head moves. We therefore use the normalisation procedure that proved most robust across the naturalistic broadcast data, while recognising that no approach eliminates these sources of noise completely.

Despite these limitations, the pipeline provides a scalable and reproducible means of detecting and measuring gesture across datasets that would be impractical to analyse manually, while also identifying movements that human annotators may overlook because of attention limitations or annotation fatigue.

References

- Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. 2019. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):172–186.
- Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S. Huang, and Lei Zhang. 2020. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *CVPR*.
- Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. [Learning phrase representations using RNN encoder-decoder for statistical machine translation](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.
- Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. 2022. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Gaëlle Ferré. 2018. [Gesture/speech integration in the perception of prosodic emphasis](#). In *Speech Prosody 2018*, pages 35–39.
- Isabella Fritz, Sotaro Kita, Jeannette Littlemore, and Andrea Krott. 2021. [Multimodal language processing: How preceding discourse constrains gesture interpretation and affects gesture integration when gestures do not synchronise with semantic affiliates](#). *Journal of Memory and Language*, 117:104191.
- Zheyang Gao, Jinyan Chen, Yuxin Liu, Yucheng Jin, and Dingxiaofei Tian. 2025. A systematic survey on human pose estimation: upstream and downstream tasks, approaches, lightweight models, and prospects. *Artificial Intelligence Review*, 58(3):68.
- Andrew Hardie. 2012. [Cqpweb — combining power, flexibility and usability in a corpus analysis tool](#). *International Journal of Corpus Linguistics*, 17(3):380–409.
- Dorothea Horst, Franziska Boll, Christina Schmitt, and Cornelia Müller. 2014. Gesture as interactive expressive movement: inter-affectivity in face-to-face communication. *Body-language-communication. An international handbook on multimodality in human interaction*, 2:2112–2125.
- Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. 2024. Sapiens: Foundation for human vision models. *arXiv preprint arXiv:2408.12569*.
- Sotaro Kita. 2000. *How representational gestures help speaking*, page 162–185. Language Culture and Cognition. Cambridge University Press.
- Sotaro Kita, Martha W. Alibali, and Mingyuan Chu. 2017. [How do gestures influence thinking and speaking? the gesture-for-conceptualization hypothesis](#). *Psychological Review*, 124(3):245–266.
- Kasper Kok, Kirsten Bergmann, Alan Cienki, and Stefan Kopp. 2016. [Mapping out the multifunctionality of speakers’ gestures](#). *Gesture*, 15(1):37–59.
- Okan Köpüklü, Maja Taseska, and Gerhard Rigoll. 2021. How to design a three-stage architecture for audio-visual active speaker detection in the wild. *arXiv preprint arXiv:2106.03932*.
- Emiel J. Krahmer and Marc G. J. Swerts. 2007. [The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception](#). *Journal of Memory and Language*, 57(3):396–414.
- Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. 2021. OpenPifPaf: Composite Fields for Semantic Keypoint Detection and Spatio-Temporal Association. *IEEE Transactions on Intelligent Transportation Systems*, pages 1–14.
- Mart Lubbers and Francisco Torreira. 2013–2026. [pypmi-ling: a Python module for processing ELANs EAF and Praats TextGrid annotation files](#). <https://pypi.python.org/pypi/pypmi-ling>. Version 1.71.
- Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. 2019. [Mediapipe: A framework for perceiving and processing reality](#). In *Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR) 2019*.
- Kyle Min, Sourya Roy, Subarna Tripathi, Tanaya Guha, and Somdeb Majumdar. 2022. Learning long-term spatial-temporal graphs for active speaker detection. In *European Conference on Computer Vision*, pages 371–387. Springer.
- Siti Nur Aisyah Mohd Robi, Muhammad Atiff Zakwan Mohd Ariffin, Mohd Azri Mohd Izhar, Norulhusna Ahmad, and Hazilah Mad Kaidi. 2024. [Active speaker detection using audio, visual, and depth modalities: A survey](#). *IEEE Access*, 12:96617–96634.
- Robert Ochshorn and Max Hawkins. 2015. Gentle: a robust yet lenient forced aligner built on kald. *arXiv preprint arXiv:1508.06477*.
- Wim Pouw and James A Dixon. 2019. Quantifying gesture-speech synchrony. In *The 6th Gesture and Speech in Interaction Conference*, pages 75–80. Universitätsbibliothek Paderborn.

Marc G. J. Swerts and Emiel J. Krahmer. 2008. [Facial expressions and prosodic prominence: Effects of modality and facial area](#). *Journal of Phonetics*, 36(2):219–238.

Ruijie Tao, Zexu Pan, Rohan Kumar Das, Xinyuan Qian, Mike Zheng Shou, and Haizhou Li. 2021. Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, page 3927–3935.

Marlijn ter Bekke, Linda Drijvers, and Judith Holler. 2024. [Hand gestures have predictive potential during conversation: An investigation of the timing of gestures in relation to speech](#). *Cognitive Science*, 48(1):e13407.

Peter Uhrig. 2022. *Large-scale Multimodal Corpus Linguistics: The Big Data Turn*. Habilitation, FAU Erlangen-Nürnberg.

Anna Wilson. 2020. [It’s time to do news again](#). *Zeitschrift für Anglistik und Amerikanistik*, 68(4):379–409.

Anna Wilson. under review. Rethinking gestural direction in english future conceptualisation: Corpus study of will-future and beyond. Submitted.

Anna Wilson, Irina Pavlova, Elinor Payne, Ilya Burenko, and Peter Uhrig. 2024. [World futures through rt’s eyes: multimodal dataset and interdisciplinary methodology](#). *Frontiers in Communication*, Volume 9 - 2024.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. [ELAN: a professional framework for multimodality research](#). In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC’06)*, pages 1556–1559, Genoa, Italy. European Language Resources Association (ELRA).

A The list of verbs

The list of verbs we used in section 6.2 for our quantitative analysis:

write, wear, wash, vote, use, urge, try, treat, trade, target, survive, surprise, succeed, struggle, spend, sing, show, shock, share, serve, sell, seek, search, satisfy, rush, risk, ring, respond, respect, request, report, reply, rent, remain, release, regard, record, recommend, read, quote, provide, prove, protect, prevent, prefer, prepare, practice, please, play, pick, persuade, pay, pack, oppose, operate, offer, offend, occur, obtain, notice, need, name, move, mix, miss, mind, mention, meet, mean, match, marry, manage, make, maintain, love, lose, load, live, listen, let, lend, leave, learn, laugh, know, kiss, judge, join, issue, involve, invite, invest, introduce, insist, inform, influence, infect, indicate, include, improve, imagine, ignore.

B Table

The table is the full version of Table 2 from the main text. It shows absolute vertical amplitudes ($|\Delta y|$ divided by movement length in frames, mean of per-movement values $\pm$ one standard deviation) for *will*-future versus the two perfect constructions, by verb segment.

C Post-processing of the OpenPose keypoints

C.1 Matching the Active Speaker to a Set of Keypoints

OpenPose predicts the coordinates of keypoints for all individuals present in a given frame, without reference to who is speaking. To identify which set of predicted keypoints corresponds to the active speaker, we use the bounding box predicted by TalkNet-ASD: for each frame, we identify candidate individuals whose head position lies within this bounding box. The nose keypoint serves as a proxy for head position; when OpenPose predicts the nose with low confidence, the average position of the eyes and ears is used in its place.⁶

OpenPose frequently detects several individuals in a single frame, and more than one detected individual’s head position may fall within the TalkNet-ASD bounding box – for instance, when a second person appears in the background of the shot, or when two individuals are seated close together. Two checks are applied to select a single candidate in such cases.

First, a candidate is discarded if the ratio of their estimated head width to the bounding-box width falls outside the range $[1/3, 3]$. Head width is estimated from the inter-ear distance where both ears are reliably detected, or otherwise from the inter-eye distance scaled by a factor of 2.5; if neither is available, this check is not applied to the candidate. This filters out candidates whose apparent size is grossly inconsistent with the bounding box, such as a background head that happens to fall inside a box sized for a foreground speaker.

Second, among the remaining candidates, the individual whose head position is closest to the bounding-box centre is selected, with the horizontal and vertical components of this distance normalised by the bounding box’s width and height, respectively. The frame is treated as ambiguous –

⁶A keypoint is considered reliable if its OpenPose confidence score is at least 0.2.

Tenses	Position	Mean $ \Delta y /\text{length} \pm \text{SD}$		Welch's t	p_{Bonf}
		<i>Will</i> -future	Tense in col. 1		
<i>Will</i> -future vs Present perfect	Auxiliary	0.028 ± 0.029	0.024 ± 0.018	3.2	0.009
	Content verb	0.029 ± 0.025	0.025 ± 0.022	4.4	5.54×10^{-5}
	Both	0.038 ± 0.031	0.033 ± 0.029	5.3	6.99×10^{-7}
<i>Will</i> -future vs Past perfect	Auxiliary	0.028 ± 0.029	0.025 ± 0.020	1.6	0.709
	Content verb	0.029 ± 0.025	0.025 ± 0.025	2.7	0.042
	Both	0.038 ± 0.031	0.032 ± 0.027	4.7	2.31×10^{-5}

Table 7: p -values are Bonferroni-corrected for six comparisons.

and no individual is selected – if the second-closest candidate’s distance is less than 1.5 times the closest candidate’s distance, i.e. if the two nearest candidates are not clearly separated in their proximity to the centre.

This pipeline is subject to two distinct failure modes, which are treated separately. The first concerns OpenPose’s keypoint predictions themselves: even for a correctly identified speaker, OpenPose may fail to predict a given keypoint. This typically occurs under two conditions: when a body part is moving at high speed, resulting in a blurred region rather than a clearly resolvable shape, and when a body part is occluded by another object or person. Occlusion is prevalent in the videos analysed here; a hand may be occluded by another hand or by a lower-third graphic (see Figure 1). In such cases, OpenPose returns a coordinate value of zero. These values are treated as missing data and are subsequently interpolated (Section C.2).

The second failure mode concerns the matching procedure itself: either no detected individual’s head position falls within the active speaker’s bounding box, or the frame is ambiguous in the sense described above. In both cases, a NaN (Not-a-Number) value is stored in place of a coordinate. NaN values are subject to the same post-processing steps as zero-valued coordinates.

A further complication concerns the scale of the predicted coordinates. OpenPose reports keypoint locations in pixel values.⁷ As pixel coordinates are a function of camera distance and angle, the same body part may be represented at substantially different scales across scenes and tracks, even for the same speaker. This variability limits the comparability of keypoint coordinates across a video and complicates both manual and automated analysis. The normalisation procedure described in

⁷These predictions are not necessarily integers and occasionally contain small negative values, which we treat as an artefact of the prediction process.

Section C.4 is designed to address this issue.

C.2 Interpolation and Extrapolation

Missing values arising from either of the failure modes described above are addressed through interpolation and extrapolation, applied separately to each track.

Interpolation estimates missing values that occur between two valid measurements. Where a keypoint coordinate is missing but is bounded on both sides by valid (non-zero) values, the missing values are estimated by linear interpolation between these two points.

Extrapolation addresses missing values that occur at the beginning or end of a track, where no valid measurement exists on one side. In these cases, the nearest valid value is substituted for the entire run of leading or trailing zeros.

C.3 Smoothing

In addition to missing values, OpenPose keypoint predictions are subject to jitter: small frame-to-frame fluctuations in coordinate values that do not correspond to actual movement of the underlying body part.

To reduce jitter, each coordinate value is replaced with a weighted average of values within a sliding temporal window. For a given frame, the window comprises the N preceding values, the current value, and the N subsequent values, yielding a total of $2N + 1$ points. We use $N = 5$ (11 points), a value selected empirically as offering the best trade-off between removing jitter and preserving small but genuine movements.

Weights are assigned according to the Epanechnikov kernel, $K(u) = \frac{3}{4}(1 - u^2)$, evaluated at $2N + 1$ points equally spaced on the interval $[-1, 1]$. This kernel assigns maximal weight to the central frame, with weight decreasing for frames further from the centre of the window. Figure 2 presents an

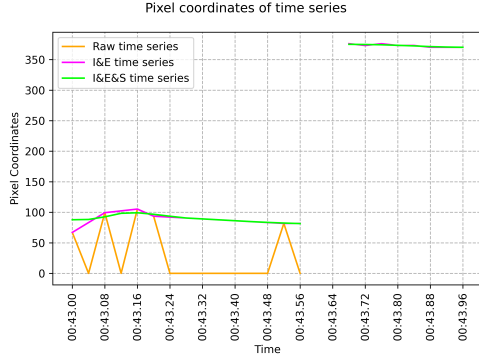


Figure 2: Raw, interpolated, and normalized and interpolated time series

example of a time series before and after smoothing.

C.4 Normalisation

Normalisation addresses the scale dependency introduced by the use of raw pixel coordinates (Section C.2). The procedure re-expresses keypoint coordinates relative to a reference point and a normalisation unit, both computed per track, yielding a coordinate space that is independent of the scale and position of the speaker within the frame.

As OpenPose predicts separate coordinate sets for the body, face, and hands, a reference point and normalisation unit are defined independently for each.

For the body, the nose coordinates serve as the reference point and are subtracted from all other body coordinates in every frame. The normalisation unit is the track-averaged distance between the nose (keypoint 0) and the neck (keypoint 1) in the BODY 25 annotation scheme. A vertical distance was chosen deliberately as the normalisation unit: TV and YouTube data frequently include scenes shot from the side, in which horizontal distances (e.g. shoulder width) are subject to foreshortening and therefore appear smaller than their true extent. Vertical distances are comparatively robust to this effect, as camera perspectives that substantially distort vertical measurements are considerably less common.

For the face, the tip of the nose serves as the reference point, and the normalisation unit is the track-averaged distance between the nasal bridge (keypoint 27) and the tip of the nose (keypoint 30) in the face keypoint annotation scheme.

Hand keypoints are not normalised, as the quality of OpenPose’s predictions for hand keypoints is

comparatively low, limiting the benefit of applying this procedure to them.

Normalisation of the body and face keypoints allows continuous time series to be produced across track and scene changes, even where the apparent scale of the speaker changes substantially, whether due to a change in shot for the same speaker or a change between speakers within the same scene.

Figure 2 illustrates the cumulative effect of interpolation and smoothing on an example time series. The orange curve represents the coordinates as originally predicted by OpenPose. The magenta curve shows the result of interpolation; extrapolation was not required in this example, as the original series did not contain leading or trailing zeros. The lime curve shows the result of smoothing applied to the interpolated (magenta) series.

Smoothing and normalisation together reduce the number of erroneously detected gestures, both in automatic annotation and in manual inspection of OpenPose-derived time series. This effect is illustrated in Figure 1, which compares raw values (top panel) with interpolated, extrapolated, smoothed, and normalised values (bottom panel) for the vertical coordinate of the left wrist. In the bottom panel, all coordinates are expressed on a common scale, without jitter or missing values.