CoordiMap: Conceptual Proposition of a new Framework for the Annotation of Verbal Elicitation Paths on Visual Experiment Stimuli and Introduction of the Associated Annotation Tool

Carmen Schacht
Ruhr-University Bochum, Germany
Department of Linguistics
carmen.schacht@ruhr-uni-bochum.de

Abstract

Consistent alignment of multi-modal experimental data—such as verbal utterances in elicitation tasks, (static) visual stimuli, and gaze data—presents a challenge in linguistic research. These elicitations often encode information about the visual perception strategies or cognitive processing of the scene. Thus, it is helpful to transform them into a structured, visually grounded format which captures the visual nature of the data, ideally able to be aligned with the corresponding gaze data. To achieve this, the present paper conceptually proposes the annotation framework for verbal elicitation paths as a data type and presents the first release of the associated newly developed *CoordiMap* annotation tool. The tool enables structured mapping of verbal elicitation data from experimental studies onto the corresponding visual stimuli. Independent of specific paradigms, the tool supports the annotation of verbal utterances in a linearized form based on coordinates directly marked on the image of the stimulus. The format is conceptually inspired by eye-tracking data formats, in which gaze behavior is represented as temporally linearized paths overlaid on the stimulus. The paper motivates the development of the tool and its annotation methodology by theoretical and experimental considerations regarding the relationship between visual perception and language production. As this a work in progress, the functionality of the annotation tool is demonstrated through an exemplary use case.

Keywords: Multi-modality, gaze data, elicitation data

1. Introduction

The collection and analysis of linguistic elicitation data in multi-modal experiments poses a challenge for researchers, particularly in terms of consistently linking diverse data modalities—such as verbal utterances, visual stimuli, and gaze data. This is especially relevant in experiments using static images as stimulus material, for example in scene or spatial description tasks, where there is a need for scientific tools that enable precise anchoring of linguistic expressions to the respective stimulus.

A typical use case involves linguistic elicitation experiments in which participants describe images containing specific visual cues or spatial configurations. These verbal data often include implicit cues about the participant's visual perception or cognitive processing of the scene (Griffin and Bock, 2000)—for example, by means of the sequence of the description or the choice of specific referential anchors (Klein, 2015). To systematically analyze such data, it is helpful to transform them into a structured, visually grounded format. The tool introduced here was developed with this goal in mind: it enables the annotation of linguistic elicitation paths directly on the stimulus image. Users can upload an image, mark relevant points according to the chosen linguistic paradigm by clicking on them, and assign a label to each point. These points are then connected in the chronological order of annotation, forming the elicitation path—a graphical representation of linear language production anchored to the visual stimulus.

The aim is to render verbal data in a form that is visually interpretable and analyzable, and that can be directly compared to other modalities—especially eye-tracking data. Just as eye-tracking operationalizes visual perception as temporally sequential paths over the stimulus, language similarly linearizes cognitive processes. The tool therefore simulates fixations, fixation durations, and saccades through the position, repetition, and connection of annotated points, offering a novel approach to analyzing linguistic and visual data within a shared coordinate framework.

The application was intentionally implemented as a framework-agnostic tool, allowing flexible integration into a wide range of theoretical and methodological research contexts and paradigms. Users can determine which linguistic units to annotate—from simple coreference expressions to complex information-structural constructions. The exported data (label, X, and Y coordinates) in `.csv` format allow for straightforward post-processing and integration into the analysis of related datasets. This tool thus provides a specialized and user-friendly platform that supports the anchoring of verbal elicitation paths on visual stimuli within a unified workflow—whether for the analysis of elicitation-only experiments or for combination with

Proceedings of The Second International Workshop onEye-Tracking Resources and Evaluation for Human-Aligned NLP, pages 35–40
May 12, 2026.

eye-tracking data based on shared stimulus materials. To promote open-access resources, *CoordiMap* will be made available under a CC BY 4.0 license[1].

2. Previous work

In empirical language research—especially within multi-modal experimental setups—a key challenge lies in linking different modalities such as language with visual attention, similarly to pragmatic metrics like gesture (Lücking et al., 2015; Pfeiffer et al., 2006) or semantic metrics such as spatial description in spatial cognition research (Delucchi Danhier, 2019). This is particularly true for experiments employing static visual stimuli designed to elicit verbal responses, which require methods for precisely mapping verbal data onto the corresponding perceived regions of an image.

A well-established method for capturing visual perception is eye-tracking, which records eye movements and yields a linearized representation of the perceptual trajectory across a stimulus. This form of data—consisting of sequences of fixations, saccades, and fixation durations—captures a temporally ordered perception path (Blake, 2013), which conceptionally aligns with linguistic production data that themselves constitute a linear representation of multidimensional cognitive processes (Ferreira and Henderson, 1998; Delucchi Danhier, 2019). Not only in spatial cognition but in all forms of language production, there is a need to abstract multidimensional information into a sequential format—a process likewise inherent to the temporally ordered nature of linguistic signals in human language (Ferreira and Henderson, 1998).

The planning and structuring of language production—known as conceptualization (Levelt, 1989)—is influenced, among other factors, by the experimental task or "Quaestio" (Delucchi Danhier, 2019). This process involves, for instance, the selection and linearization of information as well as the contextually appropriate choice of granularity in arranging that information (von Stutterheim and Carroll, 2007). Assuming that such factors are reflected not only in the composition of verbal elicitation but also in the associated gaze behavior, classic studies such as Tanenhaus et al. (1995) and Griffin and Bock (2000)—which integrated visual and linguistic data based on shared stimuli—have already demonstrated the conceptual and methodological viability of combining eye-tracking with linguistic elicitation. However, the encoding of combined data in such studies has varied, often mapping verbal data as temporal markers onto gaze paths rather than treating both as independent, structurally comparable trajectories (Griffin and Bock, 2000)—an approach that highlights the potential added value of path-based comparison proposed in this paper. This becomes particularly relevant in the case where the experiments—eye-tracking and elicitation—are performed by two separate groups of participants and thus not produce temporally matching gaze and elicitation data.

Related approaches such as 'Meaning Maps' (Henderson and Hayes, 2017, 2018) further establish visual annotation types that map image stimuli based on their semantic relevance. These mappings not only provide insights into task-driven visual salience but also serve as annotation formats for modeling attention in experimental contexts. Such methods motivate the joint investigation of gaze and language behavior in relation to shared visual stimuli—and underscore the need for an annotation tool that facilitates this type of multi-modal analysis.

2.1. Motivation for the Elicitation Path as a Data Type

Within such experimental designs, there is a need to transform verbal elicitation data into a visually grounded and formally structured data type. Verbal descriptions of image content—such as in spatial or scene descriptions—often follow implicit cognitive paths that can be traced back to the stimulus itself. Explicitly modeling these sequences as elicitation paths allows not only for the visualization of linguistic production processes but also creates a basis for direct comparison with eye-tracking data: both modalities sequentially represent perceptual and constructive processes anchored on the same stimulus. This format is particularly well-suited for experimental tasks that investigate the relationship between perception and linguistic structure—for instance, with respect to reference strategies, information structure, or spatial description patterns (Delucchi Danhier, 2019; Griffin and Bock, 2000). In contrast to individual markers that denote isolated referential points, the elicitation path enables the annotation of more complex structures that can be visualized as paths across the image—potentially aligning closely with the conceptual structure of gaze paths.

2.2. Motivation for the Annotation Tool

Existing annotation tools such as *LabelImg* (Tzutalin, 2015) or the *VGG Image Annotator (VIA)* (Dutta and Zisserman, 2019) offer robust functionality for visual image annotation (e.g., through bounding boxes), but they are not designed for the annotation of linguistic elicitation paths. However, specialized approaches like the 'Meaning Map' by Henderson and Hayes (2017), and their contribution

[1] https://osf.io/8ke4c/overview.

to linguistic research, highlight the need for task-specific tools that support linguistic annotation of visual stimuli—thereby expanding the methodological repertoire of empirical linguistics. In this context the *CoordiMap* tool was developed: a framework-agnostic, user-friendly annotation tool that makes verbal elicitation paths visually accessible on static stimuli and transforms anchors into concrete locations on the image for analysis and evaluation of the path. Users can upload a stimulus image, define relevant anchor points through simple clicks, and have these points automatically connected into paths. Each path represents a verbal utterance or cognitive sequence of linguistic production, based on the user's underlying theoretical model and can be exported as structured `.csv` files containing labels and X/Y-coordinates. The resulting data format is explicitly implemented to support comparability with eye-tracking data, as both are spatially grounded on the coordinate level and conceptually simulate fixations and saccades. This enables, for example, the investigation of whether the sequence of verbal descriptions of a scene corresponds to its visual perception—a line of inquiry particularly relevant for combined eye-tracking and elicitation studies like Griffin and Bock (2000). Importantly, the tool allows for flexible integration into a wide range of theoretical annotation frameworks—from simple coreference annotations and semantic descriptions of space (Kababgi et al., 2024; Sitter et al., 2025) to more complex information-structural annotations in the context of spatial cognition (Delucchi Danhier, 2015, 2019; Delucchi Danhier et al., 2025).

3. Functionality of *CoordiMap*

CoordiMap is a lightweight, locally run tool that is straight-forward to use and does not require an extensive amount of training. It was implemented in Python (Van Rossum and Drake, 2009) and comes with a `README` file for set-up. The tool incorporates the libraries tkinter (Lundh, 1999), pillow (Clark, 2015), numpy (Harris et al., 2020), and matplotlib (Hunter, 2007). To showcase the functionality of the new tool, the individual functions are demonstrated in a simple exemplary annotation. For demonstration purposes assume the simple annotation task of the annotation of entities represented as plain nominal mentions. A very simple exemplary elicitation regarding the example image might look like this:

(1) *'There is a tree. A bird is sitting next to a branch.'*

This elicitation thus contains the three nominal mentions of *tree*, *bird*, and *branch*, which will have to be annotated in this example. The following paragraphs will demonstrate the workflow for this annotation.

The Graphical User Interface (GUI). To run the software, the script has to be executed within the current working directory. It will then launch a GUI, where all further tasks can be carried out. Alternatively, it can be launched via the accompanying `.exe` file. The tool will be used by means of the buttons on the left side (see left side of Figure 1) and mouse clicks on the image, which can be uploaded from the user's file architecture. When a file is selected, the tool will upload it into the canvas-space in the GUI (see right side of Figure 1).

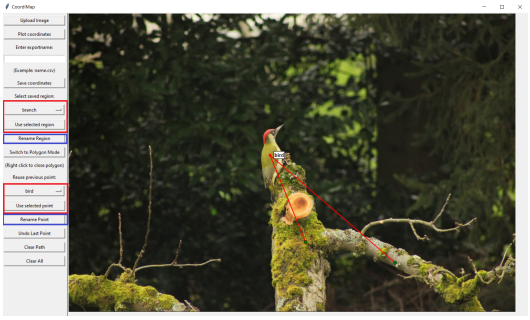

Figure 1: Exemplary annotation of the elicitation path. Reuse functionality (red) and renaming functionality (blue).

Annotation. To start annotation, the user can click the position on the image that is supposed to be annotated and thus anchor the linguistic unit to the respective position on the stimulus. The image on the canvas will then display a dot at the selected position. Every subsequent anchor is then connected to the previous one with a line creating a verbal path across the image (see Figure 1). This format is designed to mirror the format of fixations and saccades forming a view path in the data of eye-tracking experiments. To simulate prolonged fixation times typically found in eye-tracking data, the tool is able to detect multiple consecutive annotations of the same position and increase the size of the dot at this position. It will differentiate between consecutive anchors and non-consecutive anchors separated by other anchors. All individual anchors are listed in the 'point'-drop-down menu for further use. The labels are displayed next to the points on the canvas, if the user hovers the cursor over the respective point (see 'bird'-anchor in Figure 1).

Polygon Mode and Regions. In case larger entities or collective descriptions, which span extended areas of the image, need to be annotated, the tool offers a polygon mode, to create regions of adaptable sizes according to the chosen paradigm. In

the demonstration example this could apply for the mention of ‘tree’ or ‘branch’, as those span larger chunks of the image compared to ‘bird’. Those regions enable the calculation of the entities centroid–the coordinate at the center of the region–to represent the region in the form of a single anchor. To annotate regions, the mode needs to be changed from path mode to polygon mode by toggling the ‘Switch to Polygon Mode’-button. Once the mode has been changed, a region of variable size can be created by outlining the respective part of the image with clicks. The tool will automatically calculate the centroid and use it as a representative position for the region.

Figure 2: Exemplary annotation of the Polygon Mode Interface (dotted orange outline) and the Polygon Naming Interface.

Selecting and Reusing Regions and Points. To ensure a consistent annotation across participants on the same stimulus image, points and regions may be reused once annotated. This guarantees exact coordinate positions when the same unit has to be annotated. To reuse a point or a region it can simply be selected in the respective drop-down menu (see Figure 1) and used by clicking ‘Use selected point/region’.

Renaming Regions and Points. To rename a point or a region for further use, the point or region has to be selected in the respective drop-down menu. Then the user can click ‘Rename point/region’, which will open an interface, where a new label can be entered. The new label will replace all instances of the previous name of the label; even in the export file. This functionality will also update the hover-labels already displayed on the canvas (see Figure 1).

Export Coordinates of the Path. To export a finished path, a filename has to be entered into the according export interface. The export function only accepts `.csv` file-format. A `.csv`-file is then exported into the current working directory containing the annotated coordinates aligned to the stimulus in pixels as well as the label. Table 1 shows the output for the demonstration example of tree, bird and branch. The left column contains all x-coordinates of the annotations, the right column contains all the y-coordinates. After exporting the annotations, the user can either clear the current path or the entire set-up.

Label	X-Coordinate	Y-Coordinate
Tree	615	554
Bird	532	372
Branch	831	639

Table 1: Exemplary output of the export function in pixels.

4. Discussion

The presented tool constitutes a first version of a specialized annotation platform for elicitation data, aimed at precisely mapping verbal elicitation paths onto corresponding visual stimuli. Its conceptual foundation is based on the assumption that both gaze behavior and verbal description can be understood as linearizing processes of cognitive perception (Delucchi Danhier, 2019; Ferreira and Henderson, 1998). By visually anchoring linguistic units to a shared stimulus, the tool enables the identification of structural parallels between language production and perception—particularly in experimental setups that combine eye-tracking with verbal elicitation (Griffin and Bock, 2000). The tool represents an important first step toward not only conceptually but also practically linking two well-established data modalities in empirical language research: visually anchored speech and visual perception. In the long term, the framework and tool may prove useful not only in experimental linguistic contexts but also in interdisciplinary fields such as cognitive science or human-computer interaction, where the integration of multimodal data plays an increasingly central role.

In particular, the approach may prove relevant for the development and training of multimodal LLMs. The framework introduced here could support the manual data annotation processes that typically precede computational modeling, especially in contexts where aligned visual–linguistic data are required. By explicitly encoding the relationship between verbal production and visual reference points, the tool contributes to the creation of structured datasets that may facilitate the learning of grounded language representations. At the same time, it should be noted that the toy example utilized to demonstrate the functionality of the tool represents a deliberately simple application scenario, serving to illustrate the core mechanics

of the reasoning behind the scheme and the annotation process. It does, however, not exhaust the methodological potential of the approach. With the development of more complex and task-specific annotation guidelines the elicitation data could be encoded even more fully. For instance, future extensions of the framework could incorporate the temporal dimension of speech into the annotation process by integrating temporal information such as onset times, durations, or pauses. The elicitation paths as a data type could thus be enriched to reflect not only the sequential order of linguistic units but also their temporal unfolding. This would allow for an even closer comparison with eye-tracking data, in which the temporal dimension plays a fundamental role, thereby further aligning the two data types. Furthermore, the annotation framework could be expanded to explicitly incorporate the linguistic instructions of the elicitation task, since task design (the Quaestio) has a substantial impact on both linguistic production and perceptual strategies, as mentioned previously and thereby providing additional explanatory power for observed patterns in the data.

More generally, the annotation process itself could be supported by (semi-)automatic preprocessing of the elicitation data. For example, syntactic parsing, morphological tagging, or automatic coreference resolution could guide annotation decisions depending on the chosen theoretical framework. Such integrations would not only increase annotation efficiency but also form a interdisciplinary bridge between psycholinguistic research and NLP. Regarding these considerations, the current implementation should be understood as a foundational step. Its primary contribution lies in the proposition of an extensible framework that can be further adapted to complex research questions across disciplines.

5. Conclusion

The framework and tool introduced here conceptually address a methodological gap in experimental linguistics by visually anchoring verbal elicitation data and transforming it into a format that parallels the structure of eye-tracking data. The tool was developed with the goal of providing a lightweight, framework-agnostic, and locally executable instrument that can be used across a variety of experimental settings—especially for analyzing the relationship between visual perception and language production. In a follow-up step of this ongoing project, both the framework of verbal elicitation paths and the *CoordiMap* tool will be tested and evaluated empirically in a pilot annotation study comparing gaze data with elicitation paths. Future versions of the tool are intended to expand on the current functionality and adapt to the empirical demands of linguistic and cognitive research.

6. Limitations

Despite the potential of the application, the current version has several limitations:

No Automatic Alignment Functionality with Eye--tracking Data At present, the annotated elicitation paths are exported using pixel-based fixed-resolution coordinates, which are not automatically aligned with typical metrics of eye-tracking data (e.g., normalized stimulus regions, fixation durations, Areas of Interest, or the specific layout of the respective eye-tracking system). To enable direct comparability, manual post-processing—such as coordinate transformation—is currently required. Future versions may include automated coordinate alignment features to further facilitate the integration of verbal and visual paths. Additional functionality could include automatic linking of annotated points to pre-parsed linguistic features depending on the chosen paradigm—for example, syntactic or morphological information annotated in advance for each anchor point. Other annotation types, such as coreference, could also be added to mark different realizations of referential expressions.

No Empirical Evaluation with Annotated Elicitation Data As this is a work in progress, the current version of *CoordiMap* primarily serves to introduce the annotation concept and showcase its technical feasibility. A systematic application to real-world data—such as a pilot annotation study comparing elicitation and eye-tracking data—has still to be conducted. In such a future pilot study it is especially important to analyze inter-annotator consistency and conduct a usability evaluation. Only such empirical use cases will allow for a comprehensive evaluation of the tool's added methodological value for experimental linguistic research.

Theoretical Dependence of Annotation Interpretability Because the tool is designed to be framework-agnostic, the interpretability and informativeness of the annotated paths largely depend on the chosen theoretical model. The quality and granularity of the annotations may vary according to the underlying linguistic framework (e.g., coreference, information structure, semantics) and must be supported by clearly defined annotation guidelines.

Acknowledgments

We are grateful to the anonymous reviewers for their helpful comments. This research is funded by

Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102.

7. Bibliographical References

C. Blake. 2013. Eye-tracking: Grundlagen und anwendungsfelder. In W. Möhring and D. Schlütz, editors, *Handbuch standardisierte Erhebungsverfahren in der Kommunikationswissenschaft*. Wiesbaden, Springer VS. Doi:10.1007/978-3-531-18776-1_20.

A. Clark. 2015. Pillow (pil fork) documentation.

R. Delucchi Danhier. 2015. *Sprachspezifische Aspekte der Informationsverteilung in Wegan-weisungen*. Schneider Verlag Hohengehren, Baltmannsweiler.

R. Delucchi Danhier. 2019. Linearisierungsstrategien und ihr einfluss auf die informationsstruktur und die syntaktische komplexität von zimmerbeschreibungen. In Tübingen:, editor, *Raumrelationen im Deutschen*, pages 69–89. Stauffenburg.

R. Delucchi Danhier, B. Mertins, H. Mertins, and G. Schneider. 2025. Entropy as a lens: Exploring visual behavior patterns in architects. *Journal of Eye Movement Research*, 18(5).

A. Dutta and A. Zisserman. 2019. The VIA annotation software for images, audio and video. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, New York, NY, USA. ACM.

F. Ferreira and J. M. Henderson. 1998. Linearization strategies during language production. *Mem. Cognit.*, 26(1):88–96.

Z. M. Griffin and K. Bock. 2000. What the eyes say about speaking. *Psychological Science*, 11(4):274–279. PMID: 11273384.

C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R.Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. Fernandez del Rio, M. Wiebe, P. Peterson, P. Gerard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant. 2020. Array programming with NumPy. *Nature*, 585(7825):357–362.

J. M. Henderson and T. R. Hayes. 2017. Meaning-based guidance of attention in scenes as revealed by meaning maps. *Nature Human Behaviour*, 1:743–747. Doi:10.1038/s441562-017-0208-0.

J. M. Henderson and T. R. Hayes. 2018. Rehrig, g., ferreira. *F. Meaning guides attention during real-world scene description. Scientific Reports*, 8:13504.

J. D. Hunter. 2007. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95.

D. Kababgi, G. Grisot, F. Pennino, and B. Herrmann. 2024. Recognising non-named spatial entities in literary texts: a novel spatial entities classifier. In *Proceedings of the Computational Humanities Research Conference*, volume 3834 of *CEUR Workshop Proceedings*, pages 472–481.

W. Klein. 2015. *Überall und nirgendwo. Subjektive und objektive Momente in der Raumreferenz (1990)*, pages 177–207. J.B. Metzler, Stuttgart.

W. J. M. Levelt. 1989. *Speaking: From Intention to Articulation*. The MIT Press.

F. Lundh. 1999. An introduction to tkinter. *URL: www. pythonware. com/library/tkinter/introduction/index. htm*.

A. Lücking, T. Pfeiffer, and H. Rieser. 2015. Pointing and reference reconsidered. *Journal of Pragmatics*, 77:56–79. Doi:10.1016/j.pragma.2014.12.013.

T. Pfeiffer, A. Kranstedt, and A. Lücking. 2006. Sprach-gestik experimente mit iade, dem interactive augmented data explorer. In *Dritter Workshop Virtuelle und Erweiterte Realität der GI-Fachgruppe VR/AR, Koblenz*, pages 61–72.

E. Sitter, O. Momen, F. Steig, J. B. Herrmann, and S. Zarriess. 2025. Annotating spatial descriptions in literary and non-literary text. In *Proceedings of the 19th Linguistic Annotation Workshop (LAW-XIX-2025)*, pages 308–325, Vienna, Austria. Association for Computational Linguistics.

M. K. Tanenhaus, M. J. Spivey-Knowlton, K. M. Eberhard, and J. C. Sedivy. 1995. Integration of visual and linguistic information in spoken language comprehension. *Science (New York, N. Y.)*, 268(5217):1632–1634. Doi:10.1126/science.7777863.

Tzutalin. 2015. Labelimg. Free Software: MIT License.

G. Van Rossum and F. L. Drake. 2009. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA.

C. von Stutterheim and M. Carroll. 2007. Durch die grammatik fokussiert. *Zeitschrift für Literaturwissenschaft und Linguistik*, 145:35–60.