INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Reliable estimates of interpretable cue effects with Active Learning in psycholinguistic research

Marieke Einfeldt[1], Rita Sevastjanova[2], Katharina Zahner-Ritter[3], Ekaterina Kazak[4], Bettina Braun[1]

[1]University of Konstanz, Department of Linguistics, Germany
[2]University of Konstanz, Department of Computer and Information Science, Germany
[3]University of Trier, Department II, Phonetics, Germany
[4]University of Manchester, Economics/School of Social Sciences, United Kingdom

{marieke.einfeldt,rita.sevastjanova,bettina.braun}@uni-konstanz.de,
k.zahner-ritter@uni-trier.de, ekaterina.kazak@manchester.ac.uk

Abstract

Studying the relative weighting of different cues for the interpretation of a linguistic phenomenon is a core element in psycholinguistic research. This research needs to strike a balance between two things: generalisability to diverse lexical settings, which requires a high number of different lexicalisations and the investigation of a large number of different cues, which requires a high number of different test conditions. Optimizing both is impossible with classical psycholinguistic designs as this would leave the participants with too many experimental trials. Previously we showed that Active Learning (AL) systems allow to test numerous conditions (eight) and items (32) within the same experiment. As stimulus selection was informed by the system's learning mechanism, AL sped-up the labelling process. In the present study, we extend the use case to an experiment with 16 conditions, manipulated through four binary factors (the experimental setting and three prosodic cues; two levels each). Our findings show that the AL system correctly predicted the intended result pattern after twelve trials only. Hence, AL further confirmed previous findings and proved to be an efficient tool, which offers a promising solution to complex study designs in psycholinguistic research.

Index Terms: Active Learning, psycholinguistics, cue weighting estimation, stimulus selection, prosody, limited data

1. Introduction

1.1. Background

Understanding and assessing cue weights and their relation is crucial for the understanding of linguistic phenomena and studying cue weighting is therefore an essential element in experimental linguistics. Consequently, there are many of phonetic and phonological cue weighting studies in both segmental and prosodic research (e.g., [1-7]). Research designs for cue weighting studies should ideally employ a large number of orthogonally varied cues and different lexicalisations to ensure generalisability. This is often not feasible because such designs result in a very high number of experimental trials per participant. Therefore, conclusions are often drawn on limited data (small number of cues or lexicalizations).

Traditionally, studies focus on testing (a) a high number of conditions (or cues) with few lexical items (down to N = 1), (b) a high number of lexical items with few conditions or circumvent the limitation by running (c) multiple experiments. Most experiments so far were designed to focus on a detailed picture of the weighting and interplay of the tested cues and went with a high number of conditions or cues, accepting the lower generalisability in terms of lexicalisations [8-13]. Fewer studies chose to focus on the lexical generalisability and tested a high number of lexicalisations in fewer conditions, in particular in research at the prosody-pragmatics interface [14, 15]. We have previously shown that AL systems predict outcomes of a classical 2x2x2 study design (three factors with two levels each, eight test conditions) reliably and fast [16]. In the present study, we extend the application of AL systems to designs with four binary factors (16 test conditions). The reason why we use AL for our classification task is that we do not have a corpus of already labelled data at hand to train a classifier or apply statistical methods on from where we could extract cue weights. For estimating the cue weights, we need participants to label data. Since the task includes many conditions and items, we need to make this labelling process as efficient as possible. To this end we use AL.

AL is a subfield of machine learning in which learning algorithms query annotators to label hitherto unlabelled instances [17]. AL techniques have been already employed in computer science research already since the 1980s [18]. They have been applied, among others, for named-entity recognition [19], semantic parsing [20], and text classification [21]. To reduce the human effort needed to obtain an annotated corpus, AL optimises the order in which the instances are labelled by applying an appropriate sampling strategy.

Studies with AL have proven to be able to derive rules from a small set of labelled instances [16, 22, 23] and classifications can then be validated by presenting similar conditions with other items to the participants, which means it is not necessary to present each participant with every stimulus of the test set. As a result, the labelling process is sped-up compared to traditional behavioural cue weighting studies in two ways: (i) not all stimuli need to be labelled by the participants, and (ii) all cues are updated with each label, meaning that the labelling of one instance is transferred to other instances (see section 2.1.4). In addition, AL is fully data-driven since cue weights can be estimated from the participants' responses. Taken together, these properties make AL a prime candidate for linguistic cue weighting research designs.

1.2. Present study

In the present study, we compare the predictions of an AL system for two scenarios representing the outcome patterns of a 2x2x2x2 experiment (four factors with two levels each,

http://dx.doi.org/10.21437/Interspeech.2021-1524

summing up to 16 test conditions). As in [16], 16 virtual agents labelled questions as either rhetorical question (*RQ*) or information-seeking question (*ISQ*) based on predefined probabilities (see 2.1.3); stimuli were selected by the AL system and predictions for the weighting of different cue combinations were derived based on a regression-based weighting implemented in the AL system. Evaluation criteria are the reliability of the predicted probabilities and the speed with which the predictions are achieved. Reliability is evaluated based on two measures: (i) The correlation between the actual responses (labels given by the virtual agents) and the AL predictions (the higher the correlation coefficient, the better the prediction), and (ii) the root-mean-square error (RMSE) between predicted probabilities and actual responses (the lower, the better the prediction), which indicates the deviance between predicted probabilities and actual responses. The speed, with which the scenario's outcome patterns are achieved, was operationalised by analysing the relative reduction of RMS errors. Finally, we present a stopping criterion which indicates the point during the experiment where the AL predictions stabilize and there is no further improvement. The stopping criterion is computed in analogy to the speed measure (relative gain to the preceding 5 trails of the AL probabilities). It has the advantage of being able to be applied during an ongoing labelling process (in which probabilities are not know a priori), while the RMSE-based speed measurement can be calculated only post-hoc.

2. Experiment

2.1. Method

2.1.1. Scenarios

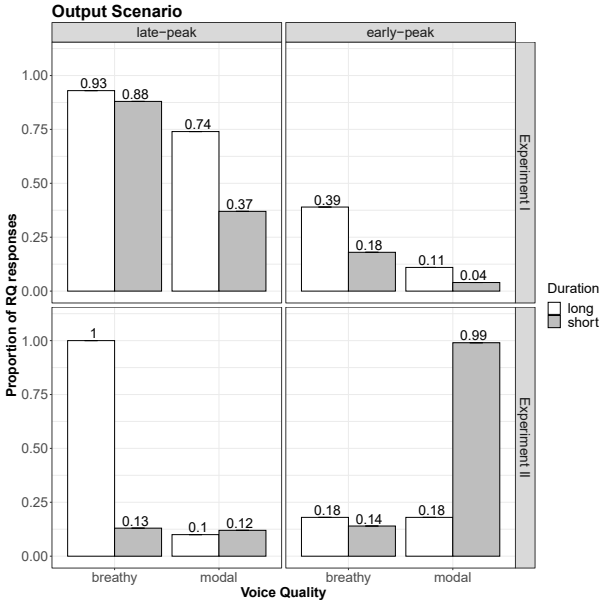

Figure 1: *A possible outcome scenario for a design with four variables, voice quality (x-axis), pitch accent (left and right facets) and duration (colors) and Experiment (upper and lower panel)*

We used the probabilities in Figure 1 for the 16 test conditions. The proportions in Experiment I were taken from the outcome of a psycholinguistic experiment [14, 16, 23] while the proportions of Experiment II were entirely hypothetical in nature:

(1) Experiment I: Stair-case pattern of the three prosodic variables (main effects)
(2) Experiment II: Interaction between all three prosodic factors

There were hence three prosodic factors (*intonation condition:* late peak vs. early peak, *duration:* short vs. long, and *voice quality:* breathy vs. modal) and one *experiment* factor (Experiment I vs. Experiment II).

2.1.2. Material

The set of stimuli consisted of 32 German *wh*-questions (e.g., *Who likes lemon?*), which were manipulated by fully crossing four factors: (a) nuclear accent type (late-peak vs. early-peak accent), (b) voice quality (breathy vs. modal voice on the final noun of the *wh*-question), and (c) duration of the utterance (lengthening or shortening of the utterance duration by 10%), resulting 64 test trails in eight test conditions (cue combinations) (see [14]) in (d) two different experimental settings (Experiment I vs. Experiment II) summing up to 16 conditions in total.

2.1.3. Virtual Agents

Virtual agents were used to simulate human participants' behaviour in a classification task in which a stimulus belongs to either the class of rhetorical question (*RQ*) or information-seeking question (*ISQ*). The main reason for the use of virtual agents was to save participant time. To this end, a binomial function was implemented, i.e., the virtual agents performed independent draws from a binomial distribution whereby the probabilities were the target probabilities from Experiment I and Experiment II (see Figure 1). As in the case of human participants, the random draws ensured some variability in responses. [16] showed that the binomial function implemented for the virtual agents replicates human responses.

2.1.4. Active Learning System

We implemented an AL system that was used to query the virtual agents for class labels. In the backend, our system iteratively learned a predictive model on the labels provided by the virtual agents, i.e., each time a new label was provided, the entire model got updated. This model played a crucial role in the AL system. It performed two tasks: (1) it predicted the class labels (*RQ* or *ISQ* class) for unlabelled stimuli, and (2) selected the next stimulus for labelling.

Regarding (1), by predicting class labels for unlabelled stimuli, we were able to reduce the number of trials needed for a labelled corpus generation. Regarding (2), we were able to obtain a stable and certain model that makes good predictions by applying an appropriate stimulus selection strategy. Strategies for stimulus selection presented in the related work are commonly classified as data-centred or model-based strategies. Data-centred strategies use the characteristics of the corpus and query labels for instances according to their similarity [24]. Model-based strategies, on the other side, integrate suggestions of a machine learning model that is trained on iteratively labelled data instances. Using different criteria, such as error reduction [25], classifier uncertainty [Smallest Margin, 24], or entropy [26], the system asks the oracle to label instances that improve the model's performance best. In contrast, data-centred strategies rely on the characteristics of the corpus and query labels for instances

according to their similarity or density [24]. In our system (following [16]), we used an uncertainty-based sampling strategy and asked the virtual agents to label stimuli where the model was most uncertain in predicting the correct label (i.e., where the probability for the prediction is close to 50% chance).

The underlying model for prediction-making can be of various complexity, ranging from simple Association Rule Mining models or linear regression models to more complex Support Vector Machines. Since the model was created iteratively, whereby the amount of data in the early stage of labelling is limited, we were restricted to using models that perform well on extremely small sample sizes. Therefore, Deep Learning models are not suitable for the present experimental set-up. In this paper, we used a linear regression model, which allowed us to predict class labels as well as to determine the importance (i.e., weight) of cues and their combinations for the prediction task. The regression-based weighting of the AL algorithm was purely data-driven (by the virtual agents' responses) and did not require an a priori specification of cue weights. This is a particularly appealing feature if one does not know whether there will be interactions between cues or not.

The system's algorithm was the following: Each time a new stimulus was labelled, the regression model provided an updated probability for each stimulus to be a *RQ* by learning a weighting of each cue and cue combination. As described before, we used the estimated probability for two purposes: prediction making and stimulus selection. If the probability of a stimulus for the *RQ* class was > 0.5, the predicted label was *RQ*; if the probability was < 0.5, the predicted label was *ISQ*. Otherwise, the predicted label was *Other* (p = 0.5). Since the regression-based weighting had the ability to extrapolate patterns to unseen conditions, the model was able to predict class labels for unseen stimuli if at least one of their cues had been learned by the model in the preceding labelling iterations.

Regarding stimulus selection, we relied on probability values and their certainty. At the beginning of the labelling process, all stimuli were assigned 0.0 probability for *RQ* class. At first, the model queries labels for stimuli with unique, not yet observed cue combinations. Later in the labelling process, the system queried labels for stimuli for which the regression model had difficulties to make predictions. That is, the system queried a label for a stimulus with the probability value that was closest to 0.5 (i.e., the most uncertain stimulus). If multiple stimuli had the same probability value, the stimulus was selected randomly from one of those.

2.1.5. Linear regression-based weighting

A linear regression model estimated the weights (w) of different cue combinations to the probability of a question to be classified as *RQ*. Each component can have a different weight:

$$P(F_1, F_2, F_3, F_4) = w_0 + w_1F_1 + w_2F_2 + w_3F_3 + w_4F_4 + w_5F_1F_2 + w_6F_1F_3 + w_7F_1F_4 + w_8F_2F_3 + w_9F_2F_4 + w_{10}F_3F_4 + w_{11}F_1F_2F_3 + w_{12}F_1F_3F_4 + w_{13}F_1F_2F_4 + w_{14}F_2F_3F_4 + w_{15}F_1F_2F_3F_4 \quad (1)$$

where F_1, F_2, F_3, F_4 (F = factor) denote the values of accent type (F_1), voice quality (F_2), duration (F_3) and experiment (F_4) respectively (these are encoded as binary $\{0,1\}$ values and the corresponding contributions of the cue combinations $\{w_0, \dots, w_{15}\}$). The weights are inferred from the labelled data with the ordinary least squares procedure (OLS) [27]. The idea of OLS is to minimise the squared distance between the estimated probability of a question to be *RQ*, $P(F_1, F_2, F_3, F_4)$ and the actual label received from a participant, by changing the weights (w). Once the weights are calibrated based on all labelled responses available, the probability of a question with any cue combination can be computed accordingly by inserting the values of the Factors 1, 2, 3 and 4 to equation 1.

This implies that this generalised probability description is flexible enough to allow for multilevel factors. Furthermore, the estimated weights w can take on any value, including negative ones. This allows us to model opposite cue effects depending on the interaction, e.g., in Experiment II the cue effects point in opposite directions. If we encode $(F_1, F_2, F_3, F_4) = (1,1,1,1)$ for a question which has an early peak, modal voice quality and short duration, the model will assign a positive value for $w_{15} > 0$ and a negative value for w_5. This assures that the contribution of short duration is positive only for this specific cue combination $(F_1, F_2, F_3, F_4) = (1,1,1,1)$ and is negative for the other conditions. Such model flexibility comes with the price of being unstable for smaller numbers of trials. It has been argued in the literature that the best option is to truncate predicted probabilities that lie outside the [0,1] interval with zero and one respectively ([28] and references therein).

2.2. Results

In this section, we first discuss the reliability of the prediction and then turn to the speed with which a reliable prediction was achieved. We then present findings on a potential stopping criterion.

2.2.1. Reliability of prediction

To assess the reliability of AL predictions, we correlated AL predictions and actual responses at different points in the labelling process (after trial 8, 16, 32, 48 and 64). Since correlation coefficients can be high despite large deviances, we further extracted RMSE, see Table 1. The correlation coefficients are high throughout, and the RMSE scores indicate a low deviation (all rs ≥ 0.83, all ps < 0.001, all RMSEs ≤ 0.22, see Table 1), suggesting that the AL predictions are reliable throughout the whole experiment.

Table 1: *Correlation analysis and RMSE values after trial 8, 16, 32, 48, 64 (df=14).*

	Scenario with four factors
Results after Trial 8	r = 0.83, p < 0.001, RMSE = 0.22
Results after Trial 16	r = 1.00, p < 0.001, RMSE = 0.01
Results after Trial 32	r = 1.00, p < 0.001, RMSE = 0.04
Results after Trial 48	r = 1.00, p < 0.001, RMSE = 0.03
Results after Trial 64	r = 0.99, p < 0.001, RMSE = 0.06

2.2.2. Speed of prediction

To assess the speed, we calculated and displayed the evolution of proportional RMSE gain (see Figure 2 for the first 30 trials) and defined a gain of under 5% compared to the average of the preceding five trials as marginal. Results showed that the marginal proportional difference to the error of the average of the five preceding trials is reached already after 12 trials.

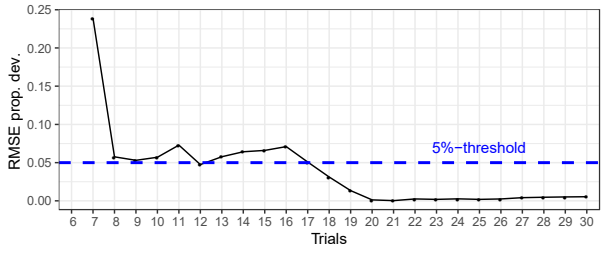

Figure 2: *Evolution of proportional differences in RMSE to previous five trails in the range to 30 trials; values form trial 31 onwards stayed at an almost identical level and are not shown.*

2.2.3. Stopping criterion

In order to be able to stop the labelling process of a participant if the AL predictions stabilize, a stopping criterion can be implemented in the system. To this end, we compared the absolute values of the predicted AL probabilities to the mean of the preceding five trials. Results show stable AL predictions at trial 17 when the gain is marginal, i.e., under 5%.

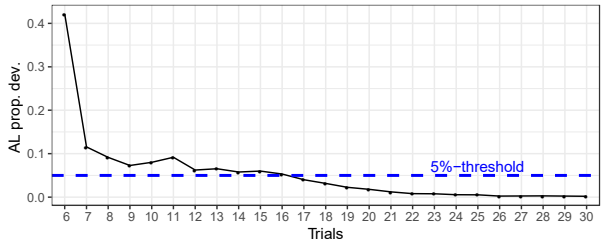

Figure 3: *Evolution of proportional differences in AL probabilities to previous five trails in the range to 30 trials; values form trial 31 onwards did not change and are not shown.*

3. Discussion

In the present study, we set out to test whether AL systems can predict outcome patterns with four binary factors (16 test conditions), i.e., more conditions than used in a previous test case [16]. Our results show that AL systems can reliably predict predetermined proportions at an early point during the labelling process, i.e., the stimulus selection informed by the regression-based AL model is reliable and fast.

The performance of the AL model was evaluated by two measures for the evaluation of the reliability (correlation coefficient and RMSE) and one for the speed (relative improvement of the proportional gain of the mean prediction RMSE compared to the average of the preceding five trials). The AL system was able to predict the outcome reliably and fast without having an already labelled dataset at hand, which serves as baseline for the cue weights and without knowing the true label of the stimulus. The latter fact is essential for the applicability of AL in linguistic cue weighting research, since knowing the weight of a cue, or a cue combination for a certain phenomenon is the goal of these studies. We showed that reliable predictions were achieved in a more complex design (four orthogonally-crossed factors, 2x2x2x2) at an early point during the labelling. Compared to experimental scenarios with eight test conditions, as presented in [16], the present 16-condition-setting is slightly slower and has lower correlation and RMSE values during the early stages (cf. [16] speed: marginal at trial 8 latest; correlation after trial 8: $r \geq 0.95$; RMSE after trial 8: ≤ 0.19). The proposed stopping criterion (based on the proportional difference of the AL probabilities compared to the average of the previous five trials), further suggests a similar point of stabilization as the speed measure. It appears to be an adequate replacement for the RMSE-based speed measurement during an ongoing labelling process. Again, the AL system in [16] reached the marginal gain faster. Yet, in spite of the slightly poorer performance in reliability and speed measures for the four-factor design as compared to the three-factor design in [16], our results in the present setting still show a reliable and quick replication of the target pattern. Hence, AL also proves to be effective in more complex study designs.

Using a linear-regression-based weighting in our AL model had the following advantages over other weighting systems: computational speed, and flexibility in modelling the direction of cue interaction effects. The proposed combination of AL and regression-based approaches has a very promising future in larger multi-level cue experiments. By design, the AL systems recalculate the predicted classification probabilities for all question with a single participant label. Such increase in effective sample size allows for an implementation of logistic regression and support vector machines to further improve on classification accuracy [29].

We are currently working on an even more complex experimental setting, in which, unlike in our test case, participants will not be able to label all target sentences due to the even higher number. A possible solution to such problems might be AL systems with a *global* learning feature and logistic regression-based weighting, i.e., stimulus selection is based on the already labelled data from the previous participants and the system further updates the cues' weights with every new item labelled across the whole labelling process. A further extension to our model would be to learn several global models each representing a separate participant group. This could account for individual differences in cue weighting.

4. Conclusions

AL systems can help to facilitate research endeavours, which aim to provide a generalizability to lexical items as well as to cues/cue combinations. Results suggest that the employed measures are dependent on the number of factors used in an experiment. A higher number of factors results in slightly lower performance of the AL during the early stages of labelling, but still produces reliable results, which are obtained at a higher speed compared to classical behavioural studies.

5. Acknowledgements

We thank the audience of the satellite workshop "Cue weighting: Thinking outside the box" at LabPhon 2020 for discussion of parts of the results. We also thank Mennatallah El-Assady for feedback on the Active Learning system, and to Nicole Dehé for comments and discussion.

The work presented here was funded by the DFG as part of research unit 'Questions at the Interfaces' (FOR 2111, project P6 and P8), grant numbers BR 3428/4-1/2 (Bettina Braun), DE 876/3-1/2 (Nicole Dehé) and KE 740/17-2 (Daniel Keim).

6. References

[1] J. Schertz and E.J. Clare, "*Phonetic cue weighting in perception and production".* WIREs Cognitive Science, 11(2): p. e1521. 2020.

[2] J. Benkí, "*Place of articulation and first formant transition pattern both affect perception of voicing in English".* Journal of Phonetics, 29: pp. 1–22. 2001.

[3] K.R. Kluender, "*Effects of first formant onset properties on voicing judgments result from processes not specific to humans".* The Journal of the Acoust Society of Amerika, 90(1): pp. 83–96. 1991.

[4] A. Gollrad, E. Sommerfeld, and F. Kügler. "*Prosodic cue weighting in disambiguation: Case ambiguity in German".* in *Proceedings of the International Conference on Speech Prosody*. Chicago, 2010.

[5] K. Kohler, "*The perception of prominence patterns".* Phonetica, 65: pp. 257–269. 2008.

[6] D.B. Fry, "*Experiments in the perception of stress".* Language and Speech, 1(2): pp. 126–152. 1958.

[7] C. Petrone, et al., "*Prosodic boundary cues in German: Evidence from the production and perception of bracketed lists".* Journal of phonetics, 61: pp. 71–92. 2017.

[8] K. Kohler, "*Terminal intonation patterns in single-accent utterances of German: phonetics, phonology and semantics".* Arbeitsberichte des Instituts für Phonetik und Digitale Sprachverarbeitung der Universität Kiel (AIPUK), 25: pp. 115–185. 1991.

[9] S. Genzel and F. Kugler, "*Production and perception of question prosody in Akan".* Journal of the International Phonetic Association, 50(1): pp. 61–92. 2020.

[10] V.J. van Heuven and M. de Jonge, "*Spectral and Temporal Reduction as Stress Cues in Dutch".* Phonetica, 68(3): pp. 120–132. 2011.

[11] K. Kohler, "*The perception of lexical stress in German: Effects of segmental duration and vowel quality in different prosodic patterns".* Phonetica, 69: pp. 68–93. 2012.

[12] D. Frost, "*Stress and cues to relative prominence in English and French: A perceptual study".* Journal of the International Phonetic Association, 41(1): pp. 67–84. 2011.

[13] O. Niebuhr and J. Winkler. "*The relative cueing power of f0 and duration in German prominence perception".* in *18th Annual Conference of the International Speech Communication Association (Interspeech)*. Stockholm, Sweden, 2017.

[14] M. Kharaman, et al. "*The processing of prosodic cues to rhetorical question interpretation: Psycholinguistic and neurolinguistics evidence".* in *Proceedings of the 20th Annual Conference of the International Speech Communication Association (Interspeech)*. Graz, Austria, 2019.

[15] K. Zahner, S. Kutscheid, and B. Braun, "*Alignment of f0 peak in different pitch accent types affects perception of metrical stress".* Journal of Phonetics, 74: pp. 75–95. 2019.

[16] M. Einfeldt, et al., "*The use of Active Learning systems for stimulus selection and data modelling in complex behavioural study designs: Results from a feasibility check".* submitted.

[17] B. Settles, "*Active learning literature survey".* Computer Sciences Technical Report at the University of Wisconsin-Madison, 2009.

[18] D. Angluni, "*Queries and concept learning".* Machine Learning, 2(4): pp. 319–342. 1988.

[19] D. Shen, et al., "*Multi-Criteria-based Active Learning for Named Entity Recognition".* in: *Proceedings of the 42nd annual meeting of the Association for Computational Linguistics (ACL-04).* Barcelona, Spain: pp. 589–596. 2004.

[20] C. Thompson, M. Califf, and R. Mooney. "*Active Learning for Natural Language Parsing and Information Extraction".* in *Proceedings of the Sixteenth International Conference on Machine Learning* (*ICML).* Bled, Slovenia: pp. 406–414. 1999.

[21] A. McCallum and K. Nigam, "*Employing EM and Pool-Based Active Learning for Text Classification"*, in *Proceedings of the Fifteenth International Conference on Machine Learning*Morgan Kaufmann Publishers Inc.: pp. 350–358, 1998.

[22] R. Sevastjanova, et al., "*Mixed-Initiative Active Learning for Generating Linguistic Insights in Question Classification"*, in *3rd Workshop on Data Systems for Interactive Analysis (DSIA) at IEEE VIS*: Berlin, 2018.

[23] M. Einfeldt, et al., "*Active learning systems as a solution for stimulus selection and data modelling in complex behavioural study designs?"*, at *Cue weighting: thinking outside the box. LabPhon 2020 Satellite Workshop*: Vancouver (online). Abstract, 2020.

[24] Y. Wu, et al. "*Sampling strategies for active learning in personal photo retrieval "*, in *Proceedings of IEEE International Conference on Multimedia and Expo, ICME* 2006, Toronto, Ontario, Canada: pp. 529–532, 2006.

[25] B. Settles, " *Active learning. Synthesis Lectures on Artificial Intelligence and Machine Learning"*. Morgan & Claypool Publishers, 2012.

[26] D.H.V. Vendrig, J., et al. "*TREC Feature Extraction by Active Learning."*. in *Proceedings of the 11th Text Retrieval Conference (TREC).* Gaithersburg, Maryland, US, 2002.

[27] J.M. Wooldrigde, "*Introductory econometrics: A modern approach"*. Toronto, Canada :Nelson Education, 2016.

[28] B. Lee, J. Lessler, and E. Stuart, "*Weight Trimming and Propensity Score Weighting".* PloS one, 6: p. e18174. 2011.

[29] J. Friedman, T. Hastie, and R. Tibshirani, "*The elements of statistical learning"*. Vol. 1, New York: Springer series in statistics, 2001.