Applying Pattern-based Classification to Sequences of Gestures

Suzanne Aussems (s.aussems@warwick.ac.uk)
Department of Psychology, University of Warwick
University Road, Coventry CV4 7AL, United Kingdom

Mingyuan Chu (mingyuan.chu@mpi.nl)
Max Planck Institute for Psycholinguistics
PO Box 310, 6500 SH Nijmegen, the Netherlands

Sotaro Kita (s.kita@warwick.ac.uk)
Department of Psychology, University of Warwick
University Road, Coventry CV4 7AL, United Kingdom

Menno van Zaanen (mvzaanen@tilburguniversity.edu)
Tilburg center for Cognition and Communication, Tilburg University
PO Box 90153, 5000 LE Tilburg, the Netherlands

Abstract

The pattern-based sequence classification system (PBSC) identifies regularly occurring patterns in collections of sequences and uses these patterns to predict meta-information. This automated system has been proven useful in identifying patterns in written language and musical notations. To illustrate the wide applicability of this approach, we classify symbolic representations of speech-accompanying gestures produced by adults in order to predict their level of empathy. Previous research that focused on isolated gestures has shown that the frequency and salience with which individuals produce certain speech-accompanying gestures are related to empathy. The current research extends these analyses of single gestures by investigating the relationship between the frequency of multi-gesture sequences of speech-accompanying gestures and empathy. The results show that patterns found in multi-gesture sequences prove to be more useful for predicting empathy levels in adults than patterns found in single gestures. This paper thus demonstrates that sequences of gestures contain additional information compared to gestures in isolation, suggesting that empathic people structure their gestural sequences differently than less empathic people. More importantly, this study introduces PBSC as an innovative, effective method to incorporate time as an extra dimension in gestural communication, which can be extended to a wide range of sequential modalities.

Keywords: Grammatical inference; speech-accompanying gestures; empathy; pattern-based sequence classification.

Introduction

People naturally accompany their speech with gestures. Several studies have reported results indicating that gesture type, frequency, and salience are related to personality traits, cognitive abilities, and empathy levels (Hostetter & Alibali, 2006; Chu & Kita, 2011; Hostetter & Potthoff, 2012; Chu, Meyer, Foulkes, & Kita, 2014). For example, Chu et al. (2014) found that empathy (i.e., how much people think about other people's thoughts and feelings) predicts the frequency of gestures with an interactive function, that is, conduit and palm-revealing gestures. Whereas previous studies mainly looked at the frequency of isolated gestures, the current research aims to extend these analyses to sequences of gestures. Even though the frequency of gestures might be the same among people with different levels of empathy, these people may order different types of gestures, most notably interactive gestures (i.e., conduit and palm-revealing gestures) in different ways. That is, in some situations more information might be hidden in frequencies of gesture *sequences* than in frequencies of single gestures.

Previous studies on cross-linguistic differences in speech-accompanying gestures (see Kita (2009) for a review) suggested that looking at gesture sequences may be fruitful. For instance, in verb-framed languages such as Turkish and Japanese, path information is expressed in one clause and manner information in another clause, in contrast to English, in which manner and path are expressed in a single clause. The verb "rolling down" is expressed in one clause in English, but it takes two clauses (e.g., "rolling/spinning" and "descending/downwards path") to express this verb in Turkish and Japanese. Research has shown that such linguistic structures influence the ways in which gestural communication is structured (Özyürek & Kita, 1999; Kita & Özyürek, 2003). Whereas Turkish and Japanese speakers tend to use one gesture to depict the rolling movement, and one gesture to depict its downward path, English speakers tend to depict manner and path in a single gesture. These studies suggest that in some languages, multi-gesture sequences depict one event, and accordingly, that the order in which people produce gestures alongside their speech may follow particular patterns.

There may be other sequential regularities present in speech-accompanying gestures. For example, different types of gestures represent different types of information in narrative, and these gestures may be ordered in a systematic way. Representational gestures often accompany speech with "narrative-level" information, which is about events and situations in the story (e.g, "A cat is looking at a canary bird in a cage."). Beat gestures often accompany speech with "meta-narrative-level" information (McNeill, 1992) which refers to the structure of the story (e.g., "The cat tries to catch the canary bird in different ways, but he never succeeds."). Interactive gestures (Bavelas, Chovil, Lawrie, & Wade, 1992) often accompany speech with "para-narrative-level" information, which refers to the interactive exchange between the speaker and the listener (e.g., "Do you know one of these American cartoons?"). These types of information may be ordered in a particular way in narrative; for example, a cluster of meta-narrative utterances may be followed by a long sequence of narrative utterances. This would in turn lead to systematic patterns in sequences of gestures. Manually annotating such regularities in gestural communication is very time consuming and inefficient, which is why it is important to investigate if such regularities can be identified automatically.

In this paper we propose a pattern-based classification approach to extend the analyses of single gestures to multi-gesture sequences. In order to demonstrate the applicability of the system, we use empathy scores as meta-information to classify the gesture sequences. Empathy may be related to the ways in which people structure information during conversation (people with high empathy levels may order information in ways that are more helpful to the listener than people with low empathy levels), which may result into particular sequences of gestures. To our knowledge, this is the first study that uses a pattern-based learning system to identify regularly produced sequences of speech-accompanying gestures and relates these to empathy. We hypothesize that multi-gesture patterns predict empathy levels in adults better than information extracted from isolated gestures.

Our classification approach is based on an existing pattern-based learning system (van Zaanen & Gaustad, 2010). This system has proven to be useful in identifying patterns in several sequential modalities, including semantics in written language (van Zaanen & van de Loo, 2012) and musical notations (van Zaanen, Gaustad, & Feijen, 2011). Our main aim is to demonstrate the wide applicability of the PBSC system. In this paper, we show the effectiveness of the system in the context of gestural communication.

The methodology we use is similar to that of Schmid, Siebers, Seuß, Kunz, and Lautenbacher (2012), who use a pattern-based sequence classifier to predict pain levels with patterns in action units that describe facial expressions. Their approach requires manual tuning of the learned patterns and can only make a distinction between two classes (pain or no pain). In contrast, our system can be applied without manual intervention and can make distinctions between any preselected number of classes, which correspond to different levels of empathy in the current study. Classifying into only two classes (high and low empathy) may be insufficient, because it leads to greater variability in people's empathy levels within a class. Our system's ability to increase the number of classes results in more specific information about the empathy level of a person, which utilizes the uniqueness in gesture sequences that people produce. We proceed by describing our system in more detail in the next section.

Pattern-Based Sequence Classification

Pattern-based sequence classification (henceforth PBSC) is an approach that aims to identify patterns in longer sequences of symbols. The patterns describe regularities found in sequences that come from the same class. Given a sequence, PBSC uses the identified patterns to assign the sequence to the class it belongs to. This approach stems from the field of grammatical inference, which addresses the task of building a compact representation of a class given a subset of sample sequences from that class (van Zaanen & Gaustad, 2010). In contrast to other grammatical inference systems, PBSC aims to learn a representation that describes the boundaries between multiple classes (corresponding to the number of empathy levels in the current study), allowing for the classification of sequences into their corresponding class. This is done by extracting patterns in the shape of sub-sequences, i.e., consecutive symbols, from the sequences in the training dataset. For practical purposes, patterns have a predetermined, fixed length (although combinations of different pattern lengths are possible as well) which coincides with the notion of n-grams, where n defines the length of the pattern (Heaps, 1978). The system only retains and uses patterns that are deemed useful according to some "usefulness" measure or scoring metric. A sequence can then be classified into a class based on which patterns are found in the sequence and the scores of the matching patterns.

System Walk-Through

PBSC, like other supervised classification systems, involves a training and a testing phase. During training, the system receives a collection of sequences that are labeled with its underlying class. First, from these sequences, all possible n-gram patterns (n consecutive symbols) are extracted and for each pattern the scoring metric is calculated indicating how well the pattern fits each of the possible classes. This results in a set of patterns with a score for each class. These patterns can be seen as vectors in a multi-dimensional space with one dimension per class. Summing pattern vectors for each occurrence in a sequence results in a vector that describes the sequence in the vector space. Second, based on the patterns, all training sequences are inserted in the vector space (and their correct class is known).

During testing, the system needs to assign a class to a new, unseen sequence. First, the system builds a vector for the sequence using the patterns. Next, it identifies the vector (of the training sequences) that has the lowest cosine distance to the vector of the test sequence. The class belonging to the training vector is returned. This corresponds to a k-nearest neighbor approach (Cover & Hart, 1967) with $k = 1$.

Scoring Metric

During training, the system aims to identify patterns that are maximally discriminative between classes. Patterns that occur frequently in a class are assigned a high score for that class compared to patterns occurring less frequently in that class, because frequent patterns describe sequences from that class better than less frequent patterns. Additionally, patterns that occur only in sequences in a particular class are more discriminative compared to sequences occurring in all classes. The combination of these properties are described in a well-known scoring metric taken from the field of information retrieval: *tf*idf* (Sparck Jones, 1972). This measure, which is extended to handle patterns (van Zaanen & Gaustad, 2010), consists of two components: term frequency (*tf*) which measures the relative frequency of the pattern and inverse document frequency (*idf*) which measures the discriminative power of the pattern over all classes. The *tf* is defined as the relative frequency of the pattern with respect to the total number of patterns found in the sequences belonging to that class. The *idf* is the logarithm of the total number of classes divided by the number of classes containing the pattern. Thus, *tf*idf* provides a score describing the discriminative power of the pattern with respect to each class.

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

$$idf_i = \log \frac{|C|}{|\{c \in C : t_i \in c\}|}$$

$$tf{*}idf_{i,j} = tf_{i,j} \times idf_i$$

Here, $n_{i,j}$ describes the number of occurrences of pattern t_i in class c_j and C denotes the set of classes under consideration.

Note that a pattern that occurs frequently in a particular class has a higher *tf* score compared to the classes in which the pattern occurs less frequently. However, the *tf* score of a pattern is weighted by the *idf* component. Patterns occurring in all classes will have a zero *idf* value, in contrast to patterns occurring in fewer classes, which will have higher *idf* values. Patterns that have a *tf*idf* score of zero for all classes (because they occur in all classes) are not retained, as they are useless for classification purposes. Note that when no matching patterns are found, the system falls back on a majority class baseline. This baseline measurement leads the system to classify a sequence into the class that occurs most frequently in the training data.

The length of the patterns has impact on the *tf*idf* scores as well as their practical usefulness in classification. In general, short patterns occur more frequently in both training and testing data. On the one hand, during training, very short patterns are likely to occur in all classes, leading to zero scores. On the other hand, it is more likely to find a short pattern (with non-zero *tf*idf* score) in the test data compared to a very long pattern (corresponding to a very specific sequence of symbols). This means that (depending on the amount of training data available), there is a sweet spot in which a specific pattern length performs best. Previous research has shown that the best results are often found with pattern lengths of three or four symbols (van Zaanen & Gaustad, 2010; van Zaanen et al., 2011).

Data

We used the dataset developed by Chu et al. (2014), which represents a sample of 122 English native speakers (71 female, 51 male) with a mean age of 19.41 years (*SD* = 4.85). This dataset contains a total of 11,032 annotated speech-accompanying gestures elicited by description tasks (for more information see Chu et al. (2014). In addition, participants were tested on several cognitive abilities and their level of empathy. Here, we focus on the relationship between the gestures participants produced alongside their speech and their level of empathy.

Empathy Quotient

In the study by Chu et al. (2014), the Empathy Quotient questionnaire (Baron-Cohen & Wheelwright, 2004) was used to measure the empathy levels of adult participants. This instrument comprises 40 questions related to empathy (e.g., "In a conversation, I tend to focus on my own thoughts than on what my listener might be thinking") and 20 filler questions unrelated to empathy ("I prefer animals to humans"). Participants were instructed to rate how strongly they agreed or disagreed with each statement (agree strongly, agree slightly, disagree slightly, or disagree strongly). On each item of the task, participants scored 2 points if the response reflected empathy strongly, 1 point if the response showed empathy slightly, or 0 points if the response did not show empathy at all. A total score was computed to indicate the level of empathy of each participant, with a maximum score of 80.

Data Representation

In the dataset, each gesture was annotated with information about its semantics, salience, and handedness. For the input of our PBSC system, we extracted this information and converted it into three distinct datasets of symbolic gesture sequences. First, the semantics of the gestures was denoted by seven unique symbols that provided information about the different types of gestures, such as representational gestures, beat gestures, and palm-revealing gestures, unclear representational gestures, representational gestures specifically used for indexing the listener, unclear gestures in general, and gestures that did not belong to the mentioned categories. Second, the level of salience of the gestures was denoted by four symbols indicating the part of the arm that was used to produce the gesture (finger, forearm, hand, or whole arm). Third, handedness was represented by three symbols that included information about whether speakers gestured with their right or left hand, or with both hands. In addition to the denotations of the latter two gesture representations, we also incorporated information (five unique symbols) about gestures that were produced with the arm only, feet, legs, torso, and head.

Classification Tasks

The PBSC system assigns participants to an empathy level class based on pattern occurrences in the (sequences of) gestures they produce. Having a partition of two classes corresponds to classifying into high or low empathy classes, whereas three classes corresponds to low, mid, or high empathy classes. To define empathy-level classes, we first divided the range of empathy scores from all participants by the number of classes to obtain the size of sub-range of empathy scores for each class, and then classified participants into the different empathy-level classes. For example, when the class size was two, participants who scored anywhere between the minimum and the minimum + (maximum – minimum) / 2 were classified into the low-empathy level class, and the rest, into the high-empathy level class. The gesture sequences produced by participants with empathy scores belonging to the same class were considered example sequences from that class. We varied the number of classes in the partition from two to six, which resulted in five classification tasks. During testing, gesture sequences produced per participant were classified and the performance of the system was measured by classification accuracy (percentage of participants classified in the correct empathy level class based on their gesture sequences). Note that it is expected that the overall system performance will decrease as the number of classes increases, because increasing the number of classes has an impact on the number of class boundaries that PBSC should learn, which makes the classification task harder. At the same time, relatively less training data is available per class when the number of classes is increased (as the participants are divided over the classes available). In contrast, the *idf* factor in the scoring metric performs better with a high number of classes (with two classes, only one non-zero *idf* value is available, with six classes, five distinct *idf* values are available).

Comparison of Results

In order to show that sequences of gestures provide more information about empathy levels than single gestures, we need to compare the performance of the PBSC system using longer patterns (n = 2, 3, 4, 5, or 6) with the performance of the PBSC system using single gestures (n = 1). Thus, our analysis includes six pattern lengths.

The accuracy of the system was measured through 10-fold cross-validation. This procedure involves randomly breaking up the dataset into ten folds of equal size and subsequently training the system based on nine of these folds to test on the tenth (unseen) fold. This process is then repeated until all ten folds have been tested once and a mean accuracy is computed for the system's performance.

Results

The accuracy of classification by the PBSC (0–100%) was entered in a 3 (gesture representation) x 5 (classification task) x 6 (pattern length) ANOVA. The results revealed no main effect of gesture representation on system performance, $F(2) = 0.251, p = .778$. Moreover, gesture representation did not significantly interact with the other two variables in the design. Thus, it did not matter if a gesture was described based on its semantics, salience, or handedness; the system performance was not affected by the symbolic representation of the gestures.

In Figure 1, horizontal lines represent the classification accuracy when the system used information extracted from single gestures (n = 1). The other lines illustrate the classification accuracy when the system used gesture sequences (n = 2, 3, 4, 5, or 6). As can be seen, increasing the number of classes to classify into (illustrated in the different panels) leads to lower accuracy scores overall, which is an artifact of the system.

The ANOVA revealed a significant interaction effect between pattern length and classification task on classification accuracy, $F(20) = 7.901, p < .001$. Tukey's HSD comparisons indicated that when the system classified participants into two or three classes of empathy, varying pattern lengths did not affect classification accuracy significantly. This is due to the fact that the *idf* has limited impact in these situations. In fact, when classifying into two classes, the system often falls back on the majority class baseline. When participants were assigned to four classes, the PBSC system that used sequences of three or more gestures to predict adults' empathy levels outperformed the PBSC system that used single gestures ($p < .001$). Additionally, the classification accuracy of the system was significantly higher when using sequences of three or more gestures than when extracting information from sequences of two gestures ($p = .009$ for n=3, $p < .001$ for all other pattern lengths). This indicates that long patterns lead to higher classification accuracy than short patterns. Pattern lengths had an effect when participants were classified into five classes: the system that used sequences of two or more gestures to predict empathy outperformed the system that used single gestures ($p < .001$). It is not surprising that a significant pattern length effect was found for these classification tasks. With a high number of classes to classify into, the *idf* weight is more useful (for all pattern lengths), allowing for a more fine-grained weighing of the corresponding *tf* score. Increasing the number of classes even more, leads to a decrease in amount of training (and testing) data per class, which is why we found no interaction effect when participants were classified into six classes of empathy. When the number of classes is higher than five, the amount of training data per class becomes too small to accurately find patterns in sequences of gestures. With the amount of data available from the dataset developed by Chu et al. (2014), the sweet spot seems to lie around four or five empathy-level classes and sequences of three or more gestures. When more data is available, we expect that a higher number of classes and longer pattern lengths lead to even better classification results.

Conclusion & Future Work

PBSC is a pattern-based classification approach, which has proven to be useful in predicting meta-information in a range of sequential modalities (e.g., written language, musical notations). To contribute to the wide applicability of the PBSC,

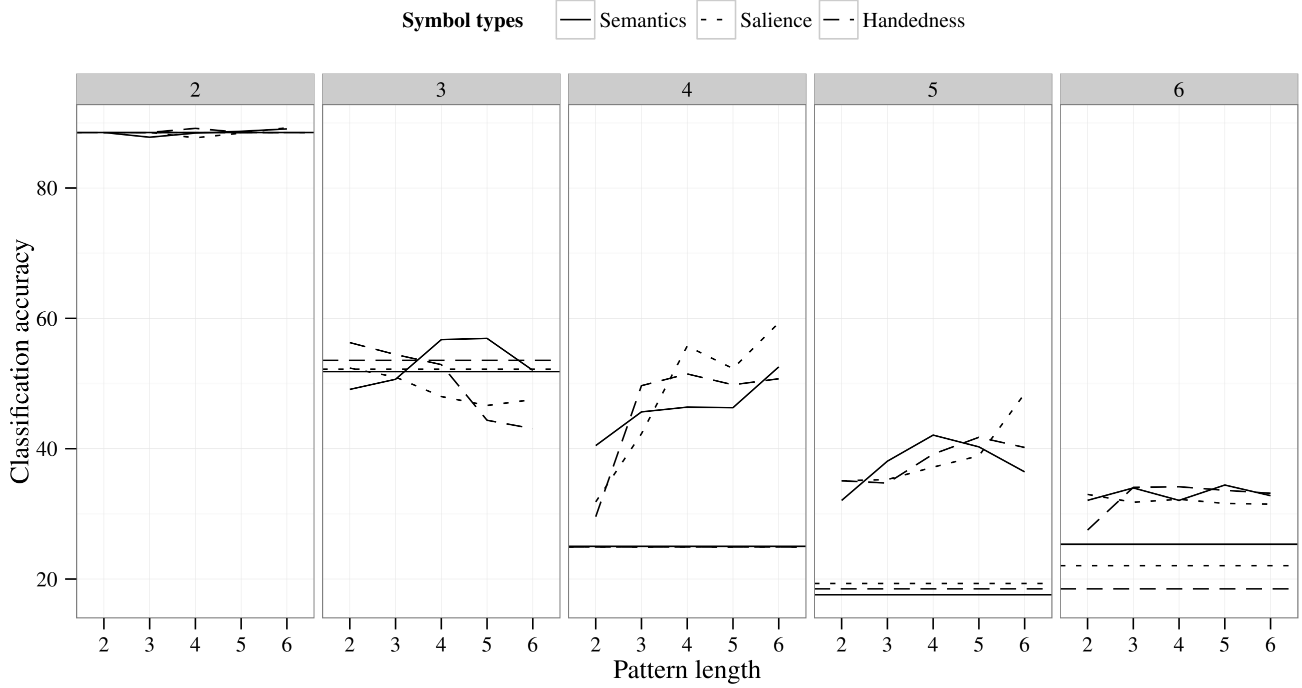

Figure 1: Accuracy of classifying participants into empathy-level classes based on their gesture sequences (y-axis), using 10-fold cross-validation. The analyses were split based on the different gesture representations (symbol types), pattern length (x-axis), and classification tasks (different panels). Horizontal lines represent classification accuracy with single gestures.

we demonstrate that the approach can also be successfully used in the context of gestural communication. As a practical example, we examined the relationship between patterns in sequences of speech-accompanying gestures produced by adults and their level of empathy.

We found that patterns describing sequences of gestures provide more discriminative power compared to patterns describing single gestures when predicting empathy levels of gesturing participants. That is, the relative frequency of multi-gesture patterns predicted participants' empathy scores better than the relative frequencies of gestures in isolation. This was the case for all three symbolic gesture representations: semantics, salience, and handedness. We found evidence for this when comparing symbol patterns consisting of one symbol with longer patterns. The differences lie within the tasks in which participants were classified into four or five empathy classes, because these classification tasks provided the system with enough training data in each class to allow for optimal discriminative power of the *idf* component of our scoring metric. This, in turn, led to more pronounced differences between the patterns. When classifying into four classes, we found additional evidence that long patterns contain more information than short patterns, as patterns of two symbols were outperformed by longer patterns. We conclude that gestures are not produced in isolation; in fact, our results indicated that they are related to each other in time. The PBSC identified this information and successfully used it to predict empathy levels in adults.

Previous research has shown that gestures are shaped in part by speakers' desire to communicate information clearly to their listeners (Hostetter, Alibali, & Schrager, 2011). Empathy levels may be related to the ways in which people structure information in conversation. Speech-accompanying gestures are related to information threads in the flow of the conversation. Speakers with a high empathy level may think more about how well the listener can follow the conversation and structure the order of information, accordingly. This may lead to specific patterns gesture sequences because different types of gestures are associated with different types of utterances (e.g., representational gestures with narrative utterances and beat gestures with meta-narrative utterances (McNeill, 1992)). Our results suggest that empathic people structure their gestural communication at the discourse level in ways that are different from less empathic people.

Several directions for future work may be considered. First, an in-depth, qualitative analysis of the patterns may be carried out to investigate, for instance, whether differences are caused by clustering of certain types of gestures at various points in narrative and/or systematicities in the use of interactive gestures alongside speech. The most discriminative patterns between the classes could provide insight into which gesture sequences are typical for a particular empathy level. Second, PBSC allows for alternative gesture representations, for instance, combining representations of different aspects of a gesture into one complex symbol. This can be used to investigate the relative importance of different aspects of gestures. Third, a cross-linguistic comparison may be interesting. Information provided in multi-pattern gesture sequences might become more pronounced in, for instance, Turkish and Japanese conversations, because in these languages certain aspects of motion events in gesture are more often sequentialized than in English. Fourth, the relationship between multi-gesture sequences and other personality traits than empathy or particular cognitive abilities can also be investigated. Finally, we believe that the PBSC approach can be applied to many other situations that deal with the classification of symbolic sequences (e.g., the visual, auditory, and motor sensory domains).

Acknowledgments

This research was supported by an Economic and Social Research Council Grant RES-062-23-2002 granted to Sotaro Kita and Antje Meyer. We thank Antje Meyer for allowing the use of the dataset. Our gratitude goes to Farzana Bhaiyat, Christina Chelioti, Dayal Dhiman, Lucy Foulkes, Rachel Furness, Alicia Griffiths, Beatrice Hannah, Sagar Jilka, Johnny King Lau, Valentina Lee, Zeshu Shao, Callie Steadman, and Laura Torney for their help with data collection and coding. We thank Birmingham City University, Bishop Vesey's Grammar School, City College Birmingham, CTC Kingshurst Academy, and University College Birmingham for their participation in our research.

References

Baron-Cohen, S., & Wheelwright, S. (2004). The Empathy Quotient: An investigation of adults with Asperger syndrome or high functioning autism, and normal sex differences. *Journal of Autism and Developmental Disorders*, *34*(2), 163-175.

Bavelas, J. B., Chovil, N., Lawrie, D. A., & Wade, A. (1992). Interactive gestures. *Discourse Processes*, *15*(4), 469-489.

Chu, M., & Kita, S. (2011). The nature of gestures' beneficial role in spatial problem solving. *Journal of Experimental Psychology: General*, *140*(1), 102-116.

Chu, M., Meyer, A., Foulkes, L., & Kita, S. (2014). Individual differences in frequency and saliency of speech-accompanying gestures: The role of cognitive abilities and empathy. *Journal of Experimental Psychology*, *143*(2), 694-709.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, *13*(1), 21-27.

Heaps, H. (1978). *Information Retrieval: Computational and Theoretical Aspects*. Orlando, FL: Academic Press.

Hostetter, A. B. (2011). When Do Gestures Communicate? A Meta-Analysis. *Psychological Bulletin*, *137*(2), 297-315.

Hostetter, A. B., & Alibali, M. W. (2006). Raise your hand if you're spatial: Relations between verbal and spatial skills and gesture production. *Gesture*, *7*(1), 73-95.

Hostetter, A. B., Alibali, M. W., & Schrager, S. M. (2011). If you don't already know, I'm certainly not going to show you! Motivation to communicate affects gesture production. In G. Stam & M. Ishino (Eds.), *Integrating gestures: The interdisciplinary nature of gesture.* Amsterdam, The Netherlands: Benjamins.

Hostetter, A. B., & Potthoff, A. L. (2012). Effects of personality and social situation on representational gesture production. *Gesture*, *12*(1), 62-83.

Kita, S. (2009). Cross-cultural variation of speech-accompanying gesture: A review. *Language and Cognitive Processes*, *24*(2), 145-167.

Kita, S., & Özyürek, A. (2003). What does cross-linguistic variation in semantic co-ordination of speech and gesture reveal? Evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and Language*, *48*(1), 16-32.

McNeill, D. (1992). *Hand and mind.* Chicago: University of Chicago Press.

Özyürek, A., & Kita, S. (1999). Expressing manner and path in English and Turkish: Differences in speech, gesture, and conceptualization. In M. Hahn & S. Stoness (Eds.), *Proceedings of the 21st Annual Conference of the Cognitive Science Society* (p. 507-512). Mahwah, NJ: Erlbaum.

Schmid, U., Siebers, M., Seuß, D., Kunz, M., & Lautenbacher, S. (2012). Applying grammar inference to identify generalized patterns of facial expressions of pain. In J. Heinz, C. de la Higuera, & T. Oates (Eds.), *Proceedings of the 11th International Conference on Grammatical Inference.* Heidelberg: Springer.

Sparck Jones, K. (1972). A statistical interpretation of term specifity and its application in retrieval. *Journal of Documentation*, *28*(1), 11-21.

van Zaanen, M. (2000). ABL:Alignment-Based Learning. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING)* (p. 961-967). Saarbrücken, Germany.

van Zaanen, M., & Gaustad, T. (2010). Grammatical Inference as Class Discrimination. In J. Sempere & P. Garcia (Eds.), *Grammatical Inference: Theoretical Results and Applications.* Berlin/Heidelberg: Springer.

van Zaanen, M., Gaustad, T., & Feijen, J. (2011). Influence of Size on Pattern-based Sequence Classification. In P. van der Putten, C. Veenman, J. Vanschoren, M. Israel, & H. Blockeel (Eds.), *Proceedings of the 20th Dutch Conference on Machine Learning* (p. 53-60). The Hague, The Netherlands.

van Zaanen, M., & van de Loo, J. (2012). Learning Interpretations Using Sequence Classification. In J. Heinz, C. de la Higuera, & T. Oates (Eds.), *Proceedings of the Eleventh International Conference on Grammatical Inference* (p. 220-223). Washington, DC.