Unsupervised acquisition of discrete grammatical categories

Shakouri, D.P.; Cremers, C.; Schiller, N.O.

UNSUPERVISED ACQUISITION OF DISCRETE GRAMMATICAL CATEGORIES

David Ph. Shakouri[1,2], **Crit Cremers**[1], and **Niels O. Schiller**[1,2,3]

[1]Leiden University Centre for Linguistics (LUCL), Leiden University
[2]Leiden Institute for Brain and Cognition (LIBC), Leiden University
[3]City University of Hong Kong (CityU), Hong Kong

ABSTRACT

This article presents experiments performed using a computational laboratory environment for language acquisition experiments. It implements a multi-agent system consisting of two agents: an adult language model and a daughter language model that aims to learn the mother language. Crucially, the daughter agent does not have access to the internal knowledge of the mother language model but only to the language exemplars the mother agent generates. These experiments illustrate how this system can be used to acquire abstract grammatical knowledge. We demonstrate how statistical analyses of patterns in the input data corresponding to grammatical categories yield discrete grammatical rules. These rules are subsequently added to the grammatical knowledge of the daughter language model. To this end, hierarchical agglomerative cluster analysis was applied to the utterances consecutively generated by the mother language model. It is argued that this procedure can be used to acquire structures resembling grammatical categories proposed by linguists for natural languages. Thus, it is established that non-trivial grammatical knowledge has been acquired. Moreover, the parameter configuration of this computational laboratory environment determined using training data generated by the mother language model is validated in a second experiment with a test set similarly resulting in the acquisition of non-trivial categories.

1 Introduction

This article presents a case study on the acquisition of discrete grammatical categories, which has been executed using a multi-agent computational laboratory environment to simulate language acquisition experiments. This system has been named MODOMA, an acronym for *moeder-dochter-machine* (Dutch for 'mother-daughter-machine'). The MODOMA implements language acquisition as the result of an interaction between an adult language model, the mother, and a daughter lan-

guage model, which is designed to acquire grammatical knowledge of the target language. The mother agent is based on the DELILAH language model generating and parsing Dutch utterances (Cremers and Hijzelendoorn, 1995/2025, cf. Cremers, Hijzelendoorn, and Reckman, 2014). During the interaction the mother agent presents sample sentences of the target language to the daughter agent. As a result, the daughter language model is defined and develops the abilitity to produce and analyze utterances corresponding to her developing grammar. Similarly to human children acquiring their mother languages, the daughter language model is able to process utterances before it has fully acquired a grammar describing the target language.

Building on the work by Shakouri, Cremers, and Schiller (2025), the goal of this study is to model the acquisition of discrete grammatical categories corresponding to parts of speech such as noun, verb or adjective. The first objective is to assess whether a statistical data analysis technique, established in the literature with respect to language produced by humans, can be applied to DELILAH's output. In particular, words can be grouped based on the observation that they occur in similar contexts, see the study by Schütze (1995) for an early example illustrating this type of approach. Considering the rapid increase in the adoption of language models, examining whether machine-generated language reflects patterns similar to those found in human language is becoming increasingly relevant. The grammatical knowledge resulting from these language acquisition simulations should be represented in such a way that it can be used by the daughter language model to generate and parse utterances. One of the key contributions of this study is that the results of this analysis are used to acquire discrete rule-based representations of grammatical knowledge encoded by graph-based feature-value pairs. In this framework, each linguistic feature such as the classification of syntactic categories is represented as a node in the graph, while the relationships between these features and their corresponding values are captured by edges. These graph structures can be used productively by the daughter language model.

In the machine learning literature, a common distinction is made between supervised and unsupervised learning. Supervised learning involves presenting the algorithm with annotated exemplars, whereas unsupervised learning processes raw unannotated data samples (cf. Duda, Hart, & Stork, 2001, pp. 16–17). Examples of supervised learning approaches to natural language processing include Data-Oriented Parsing (DOP, Bod, Scha, & Sima'an, 2003) and Memory-Based Language Processing (MBLP, Daelemans & van den Bosch, 2005). Crucially, the MODOMA only employs unsupervised learning: Children acquiring their first language do not have direct access to the grammatical knowledge of the adults providing examples of the target language. Hence, application of unsupervised learning is in accordance with the design requirements of the MODOMA. Similarly, DELILAH (Cremers & Hijzelendoorn, 1995/2025), the mother in the mother-daughter system, produces sentences, which serve as input to the language acquisition system of the daughter agent and although the mother language model employs a grammar, which assigns parses and labels to the constituents represented by the sentences she produces, these labels are not provided to the daughter. The experiments described in this article utilize clustering to model the acquisition representations of grammatical categories, namely parts of speech. In this respect, one advantage of clustering over supervised techniques such as classification is that it provides an unsupervised approach for acquiring linguistic knowledge.

Interestingly, as soon as the daughter agent has learned labels specifying her grammar, she can use this acquired knowledge to annotate previously seen and new utterances by the mother agent. In the context of the MODOMA, this acquisition technique has been called internal annotation. Internal annotation implements a form of self-supervised learning (cf. Balestriero et al., 2023, Gui et al.,

2024 for surveys and Brown et al., 2020, Devlin, Chang, Lee, and Toutanova, 2019, Lan et al., 2020, Orhan, Gupta, and Lake, 2020, Yarowsky, 1995 for some influential and notable examples). In particular, the daughter agent can employ previously acquired linguistic knowledge to acquire more complex structures. The difference between supervised learning and internal annotation is that in the case of supervised acquisition techniques the labels are provided externally with respect to the daughter, for instance by another language model such as the mother, a trained linguist, or an annotated corpus. Conversely, for internal annotation the labels are provided internally, that is, based on previously acquired grammatical knowledge by the daughter. Thus, if the annotations can be provided by the daughter, during later stages of acquisition the daughter can employ supervised acquisition techniques as well. Nonetheless, notwithstanding the use of supervised techniques the entire acquisition system is unsupervised.

2 Description of the MODOMA

The MODOMA system is the result of a research project aimed at providing a laboratory approach to language acquisition. This project builds on the DELILAH project, which resulted in the Leiden parser and generator of Dutch, a language model generating and parsing Dutch utterances such as sentences or noun phrases (NPs). The MODOMA consists of three main components: (1) DELILAH providing samples of the adult language, (2) a daughter language model, which is developed to acquire the mother language and starts communicating as soon as it has acquired any suitable construction, and (3) an interaction system enabling the conversation, that is, the exchanging of utterances between the mother and the daughter agents (cf. Shakouri et al., 2025). These three components will be briefly discussed.

DELILAH implements a combinatory list grammar, a formalism related to combinatory categorial grammar (CCG, Cremers, 2002, pp. 378-386; Cremers et al., 2014, pp. 115-137, cf. Baldridge and Kruijff, 2003 for surveys Moortgat, 1997, Steedman, 1996). Cremers et al. (2014) present detailed discussions and analyses of the design and formalisms employed by DELILAH, see also Reckman (2009, pp. 25–29)[1]. DELILAH's generator and parser execute a grammar containing graph structures representing lexical items such as Dutch words and constructions. These graphs explicitly represent grammatical properties of these words and constructions such as the phonological form, concepts, logical forms encoding meaning (cf. Cremers & Reckman, 2008), grammatical number, grammatical person, and the syntactic category (e.g., determiner, noun, or verb). This predefined grammar model is used to produce composed structures corresponding to sentences or constituents (e.g., noun phrases or verb phrases) and analyze utterances by unifying these graph structures. This type of approach to generating and parsing language is similar to the approach by Head-Driven Phrase Structure Grammar (HPSG, Pollard & Sag, 1987, 1994). DELILAH's generator takes as its input a word or concept representation and produces a utterance based on this input while for DELILAH's parser the input is a surface form string and the output is a parse providing a structural analysis of this input string. In both cases the output is represented by a template corresponding to a well-formed structure according to DELILAH's grammar. When DELILAH is called by the MODOMA, this template corresponds to a utterance, which is used to take part in the conversation with the daughter agent. For the purposes of the MODOMA project, DELILAH has been taken as is: This language model provides the samples of the adult language while the project is aimed at implementing a language acquisition system.

[1] A demonstration version presenting functionalities of this complete language model is provided at https://delilah.universiteitleiden.nl/indexen.html (Cremers & Hijzelendoorn, 2025), last accessed Jan 15, 2025.

MEMORY STACK POSITION:	0
LEXICAL ENTRY NUMBER:	1
SESSION ID:	1601581107338
CONFIDENCE LEXICAL ENTRY:	60
HEAD DIRECTIONALITY:	null
CONFIDENCE HEAD DIRECTIONALITY:	0
TERMINAL:	T
PHONFORM:	*winkelen*
SEMFORM:	A
SEMFORM INDEX:	null
GRAMMATICAL PROPERTIES:	PROPERTY TYPE: A PROPERTY VALUE: n CONFIDENCE PROPERTY: 60
HEAD:	[WINKELEN.1]
ARGUMENT:	[null]

Figure 1: Example of the representation of an acquired lexical item for *winkelen* ('shop', V) by the daughter agent

The daughter agent is a novel language model that has been specifically designed and developed for this project. It consists of three main components: (1) a generator and a parser, which enable the daughter agent to generate utterances and provide analyses as well as grammaticality judgments with respect to utterances generated by the mother language model, (2) a grammar, which is used by the generator and parser to process language and consists of templates corresponding to acquired words and grammatical constructions, and (3) a language acquisition device enabling the acquisition of knowledge of the target language and subsequently adding this newly acquired knowledge to the grammar. These components are connected to facilitate acquiring knowledge of the target language, representing it in the grammar, and using this grammar model to process utterances. Thus, language acquisition results in the construction of a functioning language model.

Similarly to DELILAH, the templates in the grammar are represented by graph structures, which explicitly specify the grammatical properties of the items. Figure 1 presents an example of a minimal template acquired by the daughter language model. These templates consist of binary graph structures containing a HEAD and an ARGUMENT position, which are designed to hold graphs of the same type as the main graph structure. Moreover, grammatical properties of the items are explicitly represented using feature-value pairs. For example, these graph structures include a phonological representation (PHONFORM), a semantic representation (SEMFORM) consisting of a unique label representing the item's denotation, an indication of whether the item is a word or a construction identified by the TERMINAL feature, the HEAD DIRECTIONALITY (i.e., currently unknown, right-headed, or left-headed) used to linearize the structure in order to generate an utterance, and the SEMFORM INDEX, which allows implementing shared referents between items to model (the acquisition of) linguistic phenomena such as reflexives and anaphora. In addition, the confidence the language model has with respect to previously acquired items and grammatical properties is represented by confidence feature-value pairs such as CONFIDENCE LEXICAL ENTRY (for the whole item) and CONFIDENCE HEAD DIRECTIONALITY (for the head directionality). Crucially, these templates enable the specification of a dynamic list of feature-value pairs to represent the results

of language acquisition procedures. This list can be employed to model the acquisition of a wide range of linguistic phenomena (cf. Shakouri et al., 2025, for experiments using this list to encode functional and content categories). As the MODOMA implements unsupervised learning, the daughter language model has no information regarding the labels used by the mother language model. Therefore, the results of language acquisition are represented by unique alpha-numeric labels such as ⟨A:n⟩ (for feature A and value *n*) for instance corresponding to ⟨PARTOFSPEECH:verb⟩.[2] In particular, the linguistic knowledge resulting from the language acquisition experiments presented by this article is saved using this list. Finally, technical metadata are similarly retained for example the combination of the SESSION ID and LEXICAL ENTRY NUMBER can be used to uniquely identify each item of acquired knowledge by the MODOMA.

Typically computational approaches to natural language processing using graph structures to represent linguistic knowledge do not make a principled distinction between the lexicon and the grammar. In more conventional grammar descriptions, the term lexicon usually refers to the acquired and retained word list and the properties of the contained words such as grammatical features, phonological form and reference (meaning) while the term grammar is reserved for the collection of abstract rules describing the language. In computational approaches employing graph structures words are usually saved using the same type of templates as rules while for rules these templates display a higher level of abstraction, that is, less or no specific words with for instance a fully defined phonological form are specified in the saved graph. Thus, for the algorithm there is no difference between what is traditionally referred to as lexicon versus grammar. For this reason, the terms grammar and lexicon are used as synonyms for the linguistic knowledge. In this respect, the MODOMA daughter language model resembles HPSG (Pollard & Sag, 1987, 1994) and the DELILAH system, which is used as the mother agent: The result of language acquisition is either a template corresponding to a word or a specification of a feature-value pair restricting the combinatory properties of a new or previously acquired template (cf. Shakouri et al., 2025).

The generator and parser of the daughter language model generate and analyze utterances by searching the grammar containing templates previously constructed by the language acquisition device and subsequently unifying graphs to create well-formed structures in accordance with the currently acquired grammar model. Similar to human children, the language learning agent should be able to employ a grammar and language it has not fully acquired yet. Therefore, from the perspective of simulating first language acquisition an important characteristic is the MODOMA daughter language model's ability to process unknown words and constructions. Crucially, the grammatical properties represented by these templates including the feature-value pairs in the dynamic list impose restrictions on which templates the generator or parser can combine, thereby limiting the set of well-formed parses and utterances of the target language. Thus, the knowledge representations enable a comprehensive language model capable of producing and analyzing utterances of the target language. Moreover, employing a language model with explicitly defined knowledge allows for the inspection of its contents. Thus, the results and development of language acquisition can be directly evaluated. Therefore, the language model is inherently an accountable artificial intelligence system, ensuring transparency and responsibility in its processes. Employing these types of models provides an interesting perspective in relation to for instance large language models.

[2]Technically, the features in the dynamic list of feature-value pairs are implemented as values of PROPERTY TYPE. However, functionally they represent features with the corresponding value of PROPERTY VALUE acting as their value.

Although DELILAH is a language model employing graph structures to represent grammatical knowledge as well, the daughter agent has no direct information on the mother agent's analyses other than what she can determine by analyzing DELILAH's output utterances. Therefore, defining abstract rules explaining noisy linguistic input data is a computationally complicated task. To this end, the language acquisition device employs a hybrid approach to language acquisition employing statistical as well as rule-based techniques to specify discrete and explicitly defined graph structures. Besides adding new items to the grammar such as learning a new word or construction, language acquisition should result in encoding grammatical properties of previously acquired items. Thus, language acquisition simulations by this system typically define a grammar model that is more specific than the previous version. In particular, these newly added grammatical properties restrict the possible combinations of templates during generation and parsing: As there is more knowlege of the target language, less utterances are considered well-formed. While previous studies demonstrated that functional and content categories can be acquired by a daughter language model (cf. Shakouri et al., 2025), the research presented in this article focuses on the hybrid acquisition of discrete categories encoding parts of speech, which can be used to subcategorize the items in the grammar.

Summarizing, while the grammar is specified as a result of language acquisition, the properties of the parser and generator are not changed during language acquisition. Thus, language acquisition results in two types of output (1) a lexicon/grammar and (2) a set of well-formed utterances allowed by this grammar. In this respect, the structure of the daughter language model reflects a basic computational model of natural language processing. According to this model, natural language processing involves three elements:

(1) a language: the (indefinite) set of all utterances in the target language;

(2) the grammar: a model describing this indefinite set of utterances;

(3) an automaton: the parser and generator executing the grammar to produce and evaluate (parse) the utterances corresponding to the language.

The language model that processes the set of utterances, consists of a combination of the automaton and the grammar. Conversely, as a result of language acquisition only the grammar model is defined and updated by the language acquisition device. The present study discusses experiments simulating the acquisition of non-trivial grammatical knowledge modelling parts of speech and utilizing this knowledge to define the grammar model.

Finally, the interaction system manages the exchange of utterances in the multi-agent system, for instance by taking care of turn taking and implementing optional feedback. This novel system enables the investigation of a wide range of theories and the execution of a diverse array of experiments. By focusing on the acquisition of discrete categories, this study aims to demonstrate the viability of the concept and lays the foundation for future investigations into the multi-agent modelling of first language acquisition. The goal is to implement a system, which can be used by researchers to test their models or assumptions regarding language acquisition by executing computational simulations. This article presents the results of such an experiment. An important advantage of providing a computational laboratory simulation of language acquisition is that this enables experiments that would be impossible or unethical to perform using human subjects. Crucially, all components of language acquisition including both agents are part of the same system. Thus, all aspects of the system can be fully configured depending on the experiment using param-

eter settings and all input, output, exchange of utterances, and processing by the system are logged and can be retrieved by researchers.

As the MODOMA system provides a computational laboratory simulation of language acquisition, many settings of the system are parametrized and user-controlled. Thus, for each experiment conducted using this system an optimal set of parameter settings has to be determined by the researchers. An example of a parameter is the number of mother utterances the daughter agent should have processed before executing the acquisition of grammatical categories. Crucially, the selected parameter settings do not imply that this specific configuration plays a part in the process of language acquisition, but rather these settings are implemented to enable simulations to assess the role of phenomena proposed to be a factor in language acquisition. Thus, the MODOMA can contribute to ongoing scientific debates on the role of properties of the input data and language acquisition process by providing a parametrized system.

Moreover, as the MODOMA constitutes an experimentation environment, all aspects of the system are recorded and can be retrieved. Examples of data that are logged, are: the parameter settings used to configure an experiment, the employed learning techniques, the utterances generated by the agents, and the acquired grammar. Thus, all factors that influence language acquisition such as the generation of samples of the target language by the mother agent, the grammar acquired by the daughter agent, and the acquisition process, are components in a single system and there are no factors that are beyond the control of the researchers. In this respect the MODOMA laboratory environment shares similarities with those used in the natural sciences as it aims to minimize factors that cannot be controlled or retrieved by the researcher. This case study builds on experiments previously conducted as part of the MODOMA project (cf. Shakouri et al., 2025) and serves as a showcase for the kind of research that can be performed using this laboratory environment. In particular, it resulted in the acquisition of non-trivial grammatical knowledge, which is represented by an explicit grammar model. This defines a language model that can be used to generate and parse sentences.

3 Related Work

To our knowledge, the combination of design properties of the MODOMA, that is, two agents in a single computational system such that one agent provides samples of a natural language while the other agent is designed to learn the target language in an unsupervised manner, is unique compared to other approaches in the field of modelling language acquisition. Crucially, this architecture provides novel research possibilities resulting in a computational language acquisition laboratory as researchers can control and measure all aspects of the simulations such as the parameter settings, input to the language acquisition procedures, executed language acquisition processes, produced utterances, and acquired linguistic knowledge by the daughter language model. This section reviews relevant literature and positions our approach within existing work, emphasizing its distinct contributions.

ter Hoeve, Kharitonov, Hupkes, and Dupoux (2022) discuss a road map to model language acquisition employing a teacher-student-loop and carried out two initial experiments related to the first steps. The first experiment focuses on modelling distinct domains using two completely separate vocabularies in an artificial language. The second experiment examines distinctive structures modeled through token repetitions. In these experiments, the teacher selects a fixed amount of training data from a collection of pre-constructed sentences and presents them to multiple students, which

are composed of language models. Subsequently, the student language models are tested on a set of other sentences from the artificial language. Moreover, a teacher language model is trained using a different subset of sentences of the artificial language and evaluated. The existing implementation of the framework contrasts with the MODOMA system particularly with respect to the training and test data, which in the context of the work by ter Hoeve et al. (2022) contain preconstructed sentences from an artificial language. Conversely, the MODOMA employs utterances that are representative of a natural language and are generated online by Delilah using her explicitly defined language model of Dutch.

The MODOMA models language acquisition from an ontogenetic perspective, that is, the development of an individual over time. From a phylogenetic perspective, research has been conducted on emergent communication utilizing multi-agent systems. An important example is the Talking heads project by Steels (2000, 2015). The Talking heads project is primarily aimed at providing a computational model of the evolutionary origins and development of language. This model employs multiple agents, which take part in language games (cf. Steels, 2001). During a language game, two agents engage in a dialogue typically making reference to their surroundings and as a result of this interaction a language is developed. An example of such a language game is the naming game (cf. Steels & Loetzch, 2012), during which words referring to objects are coined and can spread to the linguistic knowledge of other agents not involved in the initial language game that gave rise to this term. Some crucial differences when compared to the MODOMA project is that both agents acquire the target language. This contrasts with the MODOMA experimental design, in which one agent is the adult not acquiring any novel linguistic knowledge and only one agent is updating her linguistic knowledge. Moreover, most experiments conducted within the Talking heads framework employ an artificial language created by the agents themselves as a result of the language games while the adult agent in the MODOMA experimental set-up outputs utterances based on a grammar model of a natural language, Dutch. Interestingly, while initial studies employing the Talking heads framework focused on the evolutionary emergence of words and lexical development (e.g., Steels & Kaplan, 2001), more recent investigations also addressed syntactic acquisition and development (cf. Steels & Beuls, 2017; Steels, van Eecke, & Beuls, 2018; van Trijp, 2016).

Moreover, recent work by Chaabouni, Kharitonov, Bouchacourt, Dupoux, and Baroni (2020) investigates the emergence of artificial language by two deep neural agents involving different architectures such as two gated recurrent units (GRU) and a feed forward network (FFN). In particular, learning speed and generalization accuracy with respect to compositionality are assessed. A Receiver agent receives an input i consisting of i_{att} attributes and i_{val} possible values and constructs a message m based on this input. Subsequently, a Sender agent is inputted with message m and produces an output $\hat{\imath}$. This game has met its goals if the input of the Receiver resembles the output of the Sender, that is, $i = \hat{\imath}$. A follow-up study by Chaabouni et al. (2022) examines the effects of scaling up the 'dataset, task complexity, and population size'. The study employs a Speaker, which takes an image as its input to generate a message, and a Listener, which receives this message along with a collection of images. The Listener's task is to select the image that corresponds to the message. Interestingly, Chaabouni, Kharitonov, Dupoux, and Baroni (2021) conduct experiments simulating the emergence of discrete color-naming systems. In addition, Chaabouni, Kharitonov, Dupoux, and Baroni (2019) and Rita, Chaabouni, and Dupoux (2020) explored the emergence of Zipf' law with respect to artificial language created by two interacting agents. Similar to the MODOMA approach, their experimental design involves two interacting agents. Conversely, an important difference is that the language produced by the Receiver and Sender agents consists of

'simple emergent codes' while the MODOMA employs (a fragment of) a natural language. Another significant distinction is the focus on language emergence while the MODOMA aims at modelling first language acquisition. In this respect these approaches complement each other.

Furthermore, Griffith and Kalish (2007) model language evolution using (a population of) Bayesian agents engaged in iterated learning with artificial languages composed of utterance-meaning pairs. Other related research is the Baby SRL project by Conner, Gertner, Fisher, and Roth (2008, 2009), which provides computational models for the acquisition of semantic role labeling. There are several notable differences between this line of research and the MODOMA. For example, in the Baby SRL project samples of the target language are not generated by a language model but are instead sourced from the CHILDES corpus (MacWhinney, 2014) or consist of constructed sentences. Moreover, as this model is supervised, the language acquiring algorithms are provided with annotated information on the structure of the target language in contrast to the MODOMA.

Alishahi conducted computational experiments with algorithms applied to corpora that have been either artificially generated or produced by humans, to acquire grammatical phenomena. For example, the experiments by Alishahi and Stevenson (2008) simulate the acquisition of early argument structure, the studies by Alishahi and Chrupała (2009) focus on lexical category acquisition, and the research by Alishahi and Chrupała (2012) investigates the acquisition of lexical categories and word meaning. Interestingly, Matusevych, Alishahi, and Vogt (2013) present methods to improve artificially generated input data representing adult-child interactions, which are intended for use similarly to data from corpora such as CHILDES that include parent-child interactions. However, unlike the MODOMA experimental design, which aims to simulate all components involved in language acquisition within a system, the algorithm generating the input data is separate from the procedures executing language acquisition. Moreover, Beekhuizen, Bod, Fazly, Stevenson, and Verhagen (2014) present an article on the acquisition of grammatical constructions. Amongst others, this model differs in two interesting ways from the present study: (1) Their algorithm acquires these constructions based on representations of situations and utterances, which have been automatically generated taking into account the distributions in a corpus containing speech directed to children, but (2) unlike the MODOMA this model does not contain multiple agents.

Finally, McCoy, Frank, and Linzen (2020) have carried out interesting research on the computational modelling of natural language acquisition employing artificial neural networks. They investigate biases of NLP models with respect to hierarchical structure versus linear order by performing simulations involving simple recurrent networks (SRNs), gated recurrent units (GRUs), and long short-term memory units (LSTM). A crucial difference between the algorithms employed by this study to model language acquisition and the approach by the MODOMA relates to the levels of understanding an information-processing device defined by the seminal work of Marr (1982/2010, pp. 24-27) classifying algorithms at three levels: (1) computational theory, which is concerned with the goal, the logic, and the appropriateness of the procedure, (2) representation and algorithm, and (3) hardware implementation. In this respect, it could be argued that an artificial neural network provides an implementation at the level of hardware implementation while the MODOMA models language acquisition at the algorithmic level, which pertains to the representations of the input and output and the algorithm executing the transformation of input to output.

4 Methods

The primary research question addressed in these experiments is: If we can replicate for DELILAH's output the well-known finding that grammatical categories can be distinguished based on the linguistic contexts of words (cf. Schütze, 1995), can this finding be leveraged to let the MODOMA acquire grammatical categories and represent these categories by a rule-based grammar employing feature-value pairs? This question is explored through the following subquestions:

(1) Considering the different types of clustering analyses, which clustering approach should be selected for this acquisition procedure?

(2) What number of groups should be configured as a parameter setting to acquire grammatical categories corresponding to more usual grammar descriptions, if applicable?

(3) Is it possible to use a data-mining analysis to acquire an abstract/discrete grammar?

(4) To what extent do the found groupings correspond to those suggested by conventional grammar models?

(5) Can the results obtained from the training experiment be replicated in the context of the test data?

Taking into account previous studies performed with the MODOMA laboratory environment that indicated that only after being confronted with at least 10,000 exemplars, patterns were sufficiently revealed to the daughter to acquire discrete grammatical categories (cf. Shakouri et al., 2025), for this experiment only training and test sets containing 10,000 sentences generated by DELILAH were used. To carry out this experiment, the MODOMA employed R Statistics (R Core Team, 2020) to implement these analyses.[3] For this experiment a hierarchical agglomerative (bottom-up) clustering algorithm has been applied using the complete-link method, that is, data in a group are classified by taking into account the similarity with respect to the two most dissimilar data points. This technique has been selected based on experimentation by applying different algorithms on the training set with 10,000 utterances as the goal is to select the simplest algorithm that allows for the detection of syntactic categories. For example, for purposes of comparison experiments have been executed using divisive (top-down) clustering on the training data generated by DELILAH as well. As experiments using divisive clustering, which requires an additional processing step and may encounter issues with clusters assembling a few data points, revealed no significant differences compared to agglomerative clustering, a hierarchical bottom-up approach has been selected for the current experiment. Crucially, the use of clustering entails that the MODOMA implements an unsupervised algorithm, that is, no labels have been provided during the training phase by a linguist or another computer program such as the mother program or a tagger to indicate which classification training data have received. Thus, the groupings are only based on the characteristics of the dataset. This diverges from learning algorithms based on classification, for example Daelemans and van den Bosch (2005) provide an example of an approach to learning based on classification rather than clustering.

The input to the clustering algorithm consists of n-grams with concordances representing keywords in context. Two MODOMA parameters specify the number of context items to be included respectively before and after the target word for further analysis. Currently, by default these parameters

[3] Amongst others, the R libraries cluster (Mächler, Rousseeuw, Struyf, Hubert, & Hornik, 2019), plyr (Wickham, 2011, 2020), javaGD (Urbanek, 2012), rJava (Urbanek, 2009, 2017), stat (R Core Team, 2020), and tidyr (Wickham and Grolemund 2016, pp. 147–170; Wickham, Vaughan, and Girlich 2020) have been used for processing the data.

have been set to 2 context words before and after the key word, but depending on the experiment other settings could be used. A fragment of this type of list containing 57,733 exemplars is provided in Table 1 below. The translations and grammatical information regarding the target and context words have been provided solely for the convenience of the readers of this article, but these data have not been used as input to the MODOMA as this algorithm implements an unsupervised model of language acquisition.

The frequency of each combination of a target word and a context word is calculated considering its position relative to the target (e.g., as the second word before or directly preceding the target). The results for all target words are then combined into a single table, which lists the frequency of each target word and context word in a specific position. Words that are not found combined with a particular target word (for a position) but are attested for other target words, will still be listed for all key words in the table with a frequency of 0 for the unattested combinations. Using a MODOMA parameter, it is possible to specify that only data that occur in the dataset with a frequency above a threshold, should be included for further analyses. This table serves as input to the hierarchical agglomerative clustering algorithm and the clustering analysis performed is similar to the approach by Baayen, Feldman, and Schreuder (2006). It is used to calculate a correlation matrix containing Spearman correlations. The result is squared to consider only the relative distances between correlations while not taking into account the differences between positive and negative correlations. Finally, the resulting correlation matrix is used to create a distance matrix indicating the relative distances between the different key words based on the context words. The outcome is used as input to the algorithm, which produces a hierarchical agglomerative clustering that can be visualized using a dendrogram (see for example Figure 2 below and Online Appendix 1 provided in the ancillary files). Based on manually inspecting the results of these analyses concerning the training data, it is possible to specify a number of discrete groups such that they combine the most similar words into clusters. Then, each of these groups will be considered as corresponding to a grammatical category and represented by a feature-value pair encoded by graph structures in the daughter language model's grammar. In particular, each item in the MODOMA daughter grammar with a phonological form corresponding to an item listed in one of these groups will be specified for the feature-value pair representing the corresponding grammatical category. Since these properties encode the combinatorial possibilities of lexical items, the result of acquisition has implications for unification (or merge) procedures involved in generating and parsing novel utterances.

Finally, to validate the parameter settings a second experiment is executed with respect to the test data containing 10,000 exemplars of the target language, which were independently generated by the mother language model. For this purpose, the same parameter settings are applied, which have been determined based on the simulation with the training data. To assess whether the experiment with the test data produces outcomes similar to those from the training data, we quantitatively evaluate the correspondence between the categories acquired by the daughter language model in both sessions. To this end, we apply Fisher's exact test to assess whether the association between the training and test clusters is significant. To analyze the results in more detail, post-hoc tests are performed using Fisher's exact test with Bonferroni correction to identify which categories contribute most to the association.

#	Context item 1	Context item 2	Target item	Context item 3	Context item 4
1.			*werkt* work:PRES:3:SG	*niet* not	*elke* every:MASC/FEM
2.		*werkt* work:PRES:3:SG	*niet* not	*elke* every:MASC/FEM	*fiets* bike
3.	*werkt* work:PRES:3:SG	*niet* not	*elke* every:MASC/FEM	*fiets* bike	
4.	*niet* not	*elke* every:MASC/FEM	*fiets* bike		
5.			*opnieuw* again	*dient* serve:PRES:3:SG	*begeleiding* guidance
6.		*opnieuw* again	*dient* serve:PRES:3:SG	*begeleiding* guidance	*te* to
7.	*opnieuw* again	*dient* serve:PRES:3:SG	*begeleiding* guidance	*te* to	*zeggen* say:INF
8.	*dient* serve:PRES:3:SG	*begeleiding* guidance	*te* to	*zeggen* say:INF	*of* whether
9.	*begeleiding* guidance	*te* to	*zeggen* say:INF	*of* whether	*mooien* pretty:N:PL
10.	*te* to	*zeggen* say:INF	*of* whether	*mooien* pretty:N:PL	*ineens* suddenly
11.	*zeggen* say:INF	*of* whether	*mooien* pretty:N:PL	*ineens* suddenly	*zongen* sing:3:PL
12.	*of* whether	*mooien* pretty:N:PL	*ineens* suddenly	*zongen* sing:3:PL	
13.	*mooien* pretty:N:PL	*ineens* suddenly	*zongen* sing:3:PL		
			...		
57,731.		*gaan* go:PRES:3:PL	*zijn* his	*BMW's* BMW:PL	*lopen* walk:INF
57,732.	*gaan* go:PRES:3:PL	*zijn* his	*BMW's* BMW:PL	*lopen* walk:INF	
57,733.	*zijn* his	*BMW's* BMW:PL	*lopen* walk:INF		

Table 1: Illustration of key words in context for use by further statistical analyses

Figure 2: Structure of clustering based on 10,000 DELILAH exemplars for the data from session 1557861465468. For a full-size dendrogram presenting all data in a format that allows for detailed inspection of individual items, consult Online Appendix 1 provided in the ancillary files.

5 Results

An analysis of the initial training dataset of 10,000 sentences generated by DELILAH revealed that the resulting clusters often aligned with grammatical categories commonly described in traditional grammars. The acquisition procedure was performed using the following parameter settings: a single run, exposure to a minimum of 10,000 utterances, a context window of two words before and two words after the target word, detection of 14 target clusters, and a minimum frequency threshold of 10 occurrences. The structure of the dendrogram resulting from the acquisition of grammatical categories based on the initial 10,000 sentences is presented in Figure 2 above. Online Appendix 1, available in the ancillary files, presents a full-size version of this extensive dendrogram with the data in a format that allows for easier inspection. To uniquely record and identify all data generated by the MODOMA system, each session is assigned a unique identification number based on the number of milliseconds elapsed since January 1, 1970. This dendrogram is a visualization of the numerical data the algorithm employs to acquire grammatical categories. It visualizes the respective (sub)groupings based on the contextual similarities between words and can be used to enhance the understanding of the acquired knowledge in relation to the dataset. Crucially, both the numerical data and the dendrogram represent an intermediate analysis of all data and the relative distances between words. Hence, while this analysis provides valuable insights, it does not lead to the identification of a discrete number of subgroups containing the most similar words that could correspond to a discrete number of categories. These categories are determined by an additional algorithm provided by R, which takes as its input the calculated numerical distances between all words.

Several observations can be made based on the result of this analysis. It lies beyond the scope of this article to provide an exhaustive listing. Nonetheless, the main clusters and several notable subgroupings will be discussed. Interestingly, this dendrogram shows that the third main left split groups many words that modify a noun, which would largely be classified as adjectives in traditional grammar. For the most part, the second main group from the left consists of words that modify a verb, namely adverbs. In the middle of the dendrogram, words that specify nouns, are grouped: Especially numerals are presented here as well as determiners such as articles, demonstrative pronouns, and indefinite pronouns. Within this cluster, there is also a subgroup of personal pronouns consisting of *ik* ('I'), *jij* ('you', INFORM) and *u* ('you', FORMAL). Similarly, complementizers such as *dat* ('that'), *terwijl* ('while') and *nadat* ('after') are clustered and a subgroup that mainly consists of auxiliaries such as *had* ('had', SG) and *waren* ('were'), can be discerned. Notably, prepositions such as *met* ('with') and *in* ('in') are grouped as well. Nonetheless, some words clustered with these prepositions would be classified differently according to traditional grammar (e.g., *vergeten*, 'forgotten' and *zeurt*, 'nags'). Similarly, verbs such as *huilen* ('cry') and *bidden* ('pray') are grouped together. However, this verb cluster also includes words that are usually considered to belong to a different category (e.g., the adjectives *actieve*, 'active' and *onhandige*, 'clumsy'). In the dendrogram in Figure 2, perfect participles such as *beroofd* ('robbed') and *gemaakt* ('made') are clustered together. Finally, there is a cluster comprising various parts of speech according to more conventional grammar descriptions including nouns, verbs, participles, and complementizers. Interestingly, several subgroups in this cluster correspond to grammatical categories. For example, one subgroup contains prepositions such as *af* ('off', 'out'), *over* ('about'), *voorbij* ('past'), *bij* ('at', 'near'), and *uit* ('from') while another subgroup consists of verbs, such as *denken* ('think') and *slapen* ('sleep'). Another example is that *daar* ('there') and

hier ('here') are clustered together. This suggests that the groupings indentified by the algorithm are non-trivial when compared to more conventional grammar models.

6 Analyses

To be used to specify discrete categories, these hierarchical clusters should be divided into a number of groups. For the purposes of this experiment, 14 groups have been requested. The resulting clusters are listed in Table 3 below. For the convenience of non-Dutch speaking readers, some additional information on the meaning and grammatical properties of these lexical items is provided in Table 3 although these metadata have not been inputted into the unsupervised procedures by the language acquisition algorithm of the MODOMA. The contents of many clusters often appear to correspond closely with groupings made by more canonical accounts of grammar. For instance, cluster 13 only lists perfect participles. Interestingly, group 2 primarily consists of functional categories such as complementizers, prepositions and auxiliaries. Similarly, cluster 5 contains mainly words specifying nouns such as determiners (e.g., articles, demonstrative and indefinite pronouns, numerals, and personal pronouns). Groups 3 and 4 mostly list nouns while cluster 11 comprises nouns as well that are all but one diminutives. For the most part, clusters 1 and 12 consist of verbs and group 10 assembles some verbs and words related to *vraag* ('question'), that is, words taking a complement clause. Additionally, group 8 corresponds to sentential adverbs while groups 9 and 14 predominantly consist of adjectives. Conversely, cluster 6 alligns less with a category employed by traditional grammar descriptions as it contains for instance words that are usually classified as adjectives, adverbs, verbs, auxiliaries, and complementizers. Finally, group 7 coincides with diverse grammatical categories as well, but especially perfect participles and prepositions are represented.

Interestingly, the aim of the MODOMA is not to replicate the groupings established by conventional grammar accounts but to construct an abstracting model of the mother language. The goal of the daughter agent is to acquire at least a weakly equivalent grammar with respect to the mother language model. A weakly equivalent grammar considers the same set of sentences as grammatical as the mother agent. Weak equivalence is often contrasted with strong equivalence: A strongly equivalent grammar does not only allow for the same set of grammatical sentences but also assigns the same parses to them (cf. Chomsky, 1963). Thus, the weakly equivalent grammars are a subset of the strongly equivalent grammars and the acquisition of a strongly equivalent grammar would similarly result from successful acquisition procedures. In the case of the MODOMA, this would imply that after acquiring the target language the daughter language model can not only generate the same set of utterances as the mother language model but also assign the same parses in terms of the constituent structure to these utterances (i.e., it has acquired the same grammar except for label names).

Nonetheless, the correspondence with other descriptions of language proposed by linguists indicates that the daughter agent has acquired non-trivial knowledge of the target language. In this respect, it is noteworthy that the grammar DELILAH, that is, the mother language model in the MODOMA system, which produces the samples of language acquisition is based on, has been constructed by implementing an existing grammatical model and well-established analyses of Dutch. Therefore, while the acquired grammar for the present experiment does not result in a strongly equivalent model, this resemblance indicates the acquisition of a grammar that often assigns a similar parse with respect to the mother grammar. The acquisition of grammatical categories resembling those proposed by linguists can be seen as evidence of the soundness of the acquired

grammar in terms of the equivalence between the mother and daughter language models. In this case the results of acquisition align with models derived from (often) centuries of linguistic research. Conversely, if different or diverging categories were acquired compared to those proposed by linguists, this would not imply that acquisition has led to a grammar that does not account for the target data. After all, the daughter is not requested to produce a strongly equivalent grammar and it is possible that a weakly equivalent model that employs completely different categories, could predict the same set of sentences. If so, this alternative analysis could provide insights for the field of linguistics by proposing new models. Nonetheless, such a result would need to be further explained.

#	FEATURE	VALUE	Words in the cluster
1.	A	a	*actieve* ('active'), *bidden* ('pray'), *blijken* ('turn out'), *concurreren* ('compete'), *gaan* ('go'), *groeien* ('grow'), *heb* ('have', 1:SG), *huilen* ('cry'), *hun* ('their'), *komen* ('come'), *lachen* ('laugh'), *onhandige* ('clumsy'), *overlijden* ('decease'), *praten* ('talk'), *'s* ('of the', ARCHAIC DET)[4], *sociale* ('social'), *veranderen* ('change'), *verschijnen* ('appear'), *verschillen* ('differ'), *zwemmen* ('swim')
2.	A	b	*aan* ('to', PREP), *als* ('if'), *alsof* ('as if'), *dan* ('then', 'than'), *dat* ('that'), *door* ('by'), *er* ('there'), *had* ('had', SG), *hadden* ('had', PL), *heeft* ('has'), *hoe* ('how'), *in* ('in'), *is* ('is'), *met* ('with'), *na* ('after'), *naar* ('to', PREP), *nadat* ('after'), *niet* ('not'), *of* ('or', 'whether'), *op* ('on'), *tegen* ('against'), *ter* ('at'), *terwijl* ('while'), *tot* ('until'), *totdat* ('until'), *van* ('of'), *vergeten* ('forgotten'), *voor* ('for'), *waar* ('where'), *waarom* ('why'), *wanneer* ('when'), *waren* ('were'), *was* ('was'), *werd* ('became', SG), *werden* ('became', PL), *wordt* ('becomes'), *zeurt* ('nags'), *zoals* ('as'), *zodat* ('so that')
3.	A	c	*aansporingen* ('incitements'), *eigen* ('own'), *mogelijkheidjes* ('small possibilities'), *opdrachtjes* ('small assignments'), *ruimte* ('space'), *verwachting* ('expectation'), *zilveren* ('silver', ADJ)
4.	A	d	*aansporinkjes* ('little incitements'), *besluiten* ('decisions'), *besluitje* ('little decision'), *eis* ('demand', N), *feiten* ('facts'), *kleine* ('little'), *mededelingen* ('announcements'), *onderwijs* ('education'), *opdrachten* ('assignments'), *overtuiginkjes* ('little convictions'), *samenhangende* ('coherent'), *uitwerking* ('effect'), *verwachtinkje* ('small expectation'), *voorsteltjes* ('small proposals'), *welzijn* ('well being', N)

[4]This archaic determiner was used by DELILAH as part of the collocation *'s morgens*, 'in the morning'.

5.	A	e	*acht* (‘eight’), *alle* (‘all’), *blijkt* (‘turns out’), *de* (‘the’, MASC/FEM/PL), *deze* (‘this’, MASC/FEM), *die* (‘that’, MASC/FEM), *dit* (‘this’, NEUT), *drie* (‘three’), *een* (‘a’, ‘one’), *elk* (‘every’, NEUT), *elke* (‘every’, MASC/FEM), *enkele* (‘a few’), *geen* (‘no’), *het* (‘the’, NEUT:SG), *ieder* (‘every’, NEUT), *iedere* (‘every’, MASC/FEM), *ik* (‘I’), *je* (‘you’, INFORM), *jij* (‘you’, INFORM), *massa’s* (‘masses’), *menig* (‘many’, SG), *menige* (‘many’, PL), *negenennegentig* (‘ninetynine’), *sommige* (‘some’, PL), *te* (‘at’, ‘to’, COMP), *twaalf* (‘twelve’), *twee* (‘two’), *u* (‘you’, FORMAL), *veel* (‘a lot of’), *vier* (‘four’), *vijf* (‘five’), *vijftig* (‘fifty’), *weinig* (‘a few’), *werkt* (‘works’)
6.	A	f	*adequate* (‘adequate’), *altijd* (‘always’), *beseft* (‘realizes’), *binnenlands* (‘domestic’), *bleek* (‘seemed’), *concurrentie* (‘competition’), *daar* (‘there’), *fossiele* (‘fossil’), *gewenst* (‘desired’, PART), *hangen* (‘hang’), *hebben* (‘have’), *hebt* (‘have’, 2:SG), *hier* (‘here’), *lacht* (‘laughs’), *meer* (‘more’), *meteen* (‘at once’), *nooit* (‘never’), *omdat* (‘because’), *praat* (‘talks’), *voordat* (‘before’), *worden* (‘become’), *zelfde* (‘same’), *zijn* (‘are’, ‘be’, ‘his’)
7.	A	g	*af* (‘off’, ‘out’), *beschikking* (‘order’), *bij* (‘at’, ‘near’), *gebleken* (‘seemed’, PART), *gekomen* (‘come’, PART), *gekund* (‘able’, PAST PART of *kunnen*, ‘can’), *gemoeten* (‘had to’, PART), *gevolgd* (‘followed’, PART), *gezeten* (‘sat’, PART), *gezien* (‘seen’), *grove* (‘gross’), *morgens* (‘of the morgen’, ARCHAIC GEN), *ontkend* (‘denied’, PART), *over* (‘about’), *proberen* (‘try’), *rond* (‘round’), *terug* (‘back’), *toe* (‘to’, ‘at’), *toen* (‘then’, ‘when’), *uit* (‘from’), *voorbij* (‘past’), *voort* (‘forth’), *wereldwijde* (‘global’)
8.	A	h	*al* (‘already’), *bijna* (‘almost’), *daarna* (‘thereafter’), *daarom* (‘therefore’), *echter* (‘however’), *eerst* (‘at first’), *eigenlijk* (‘actually’), *even* (‘for a bit’), *gaarne* (‘with pleasure’), *gedeeltelijk* (‘partly’), *ginds* (‘over there’), *inderdaad* (‘indeed’), *ineens* (‘suddenly’), *misschien* (‘maybe’), *momenteel* (‘at the moment’), *natuurlijk* (‘of course’), *nog* (‘yet’), *nu* (‘now’), *ook* (‘also’), *opnieuw* (‘again’), *soms* (‘sometimes’), *steeds* (‘still’), *tegelijkertijd* (‘at the same time’), *thans* (‘at present’), *toch* (‘however’), *verder* (‘further’), *voorheen* (‘previously’), *waarschijnlijk* (‘probably’), *weer* (‘again’), *wel* (‘well’), *zo* (‘so’, ‘immediately’)

9.	A	i	*arrogante* (‘arrogant’), *behandeling* (‘treatment’), *beschikbare* (‘available’), *binnenlandse* (‘domestic’), *blije* (‘happy’), *boze* (‘angry’), *bruto* (‘gross’), *creatieve* (‘creative’), *culturele* (‘cultural’), *demografische* (‘demographic’), *directe* (‘direct’), *dramatische* (‘dramatic’), *ethische* (‘ethical’), *eventuele* (‘possible’), *financiele* (‘financial’), *functionele* (‘functional’), *fundamentele* (‘fundamental’), *gedachtetjes* (‘little thoughts’), *gehele* (‘complete’), *gewone* (‘normal’), *goede* (‘good’), *haar* (‘her’), *hele* (‘whole’), *hobbelige* (‘bumpy’), *invoering* (‘introduction’), *leerplichtige* (‘in compulsary education’), *medische* (‘medical’), *minderjarige* (‘underage’), *museale* (‘museological’), *nieuwe* (‘new’), *officiele* (‘official’), *onschatbare* (‘invaluable’), *onschuldige* (‘innocent’), *onwerkelijke* (‘surreal’), *Oost-Europese* (‘Eastern European’), *positieve* (‘positive’), *prachtige* (‘beautiful’), *psychische* (‘psychic’), *ruimtelijke* (‘spatial’), *schaarse* (‘scarce’), *sociaaleconomische* (‘socio-economic’), *speciale* (‘special’), *spectaculaire* (‘spectacular’), *stapsgewijze* (‘gradual’), *strafbare* (‘punishable’), *vaste* (‘fixed’), *verre* (‘remote’), *volledige* (‘complete’), *voortdurende* (‘ongoing’), *voortvarende* (‘vigorous’), *waardevolle* (‘valuable’), *zichtbare* (‘visible’)
10.	A	j	*begrijpen* (‘understand’), *horen* (‘hear’), *lezende* (‘reading’), *merken* (‘notice’), *vraag* (‘question’), *vraagje* (‘small question’), *vraagjes* (‘small questions’), *vragen* (‘questions’, N, ‘question’, V), *weten* (‘know’), *zien* (‘see’)
11.	A	k	*behandelingen* (‘treatments’), *behandelingetjes* (‘small treatments’), *beloftetjes* (‘little promises’), *beschrijvinkjes* (‘small descriptions’), *besluitjes* (‘small decisions’), *bestralinkjes* (‘small irradiations’), *eisjes* (‘small demands’), *impulsjes* (‘small impulses’), *kansjes* (‘small opportunities’), *operaties* (‘operations’, ‘surgeries’), *stimulansjes* (‘small incentives’), *vermogentjes* (‘small abilities’, ‘small assets’), *verwachtinkjes* (‘small expectations’)
12.	A	l	*behoren* (‘belong’), *beseffen* (‘realize’), *blijkende* (‘turning out’), *denken* (‘think’), *dienen* (‘serve’), *durven* (‘dare’), *gillen* (‘scream’), *kleven* (‘stick’), *lijken* (‘seem’), *om* (‘to’), *ontkennen* (‘deny’), *slapen* (‘sleep’), *snijden* (‘cut’), *stijgen* (‘rise’), *tevens* (‘also’), *wensen* (‘wish’), *wensende* (‘wishing’), *werken* (‘work’), *winkelen* (‘shop’), *zitten* (‘sit’), *zittende* (‘sitting’)

13.	A	m	*beroofd* (‘robbed’, PART), *betreurd* (‘regretted’, PART), *geaccepteerd* (‘accepted’, PART), *geboden* (‘offered’, PART), *gebouwd* (‘built’, PART), *gehoord* (‘heard’, PART), *geleerd* (‘learned’, PART), *gemaakt* (‘made’, PART), *gemerkt* (‘noticed’, PART), *gesproken* (‘spoken’, PART), *gevonden* (‘found’, PART), *gevraagd* (‘asked’, PART), *gezegd* (‘said’, PART), *onderstreept* (‘underlines’, ‘underlined’, PART), *vereisende* (‘demanding’), *voorkomen* (‘prevent’, ‘prevented’, PART), *zeggen* (‘say’)
14.	A	n	*internationale* (‘international’), *jonge* (‘young’), *meervoudige* (‘multiple’), *Nederlandse* (‘Dutch’), *omvangrijke* (‘extensive’), *raad* (‘advice’), *trage* (‘slow’)

Table 3: Groupings resulting from the cluster analysis of the data from session 1557861465468

Each grouping in Table 3 is assigned a feature-value pair representing a grammatical category. These feature-value pairs are subsequently used to specify all items in the grammar of the daughter language model that have a phonological representation corresponding to the words listed in this table for each category. Crucially, all results of cluster analysis are represented using the same feature, in this case A. However, depending on the cluster, the items are assigned different values, ranging from *a* to *n*. The feature (i.e., A) can be interpreted as corresponding to a major linguistic category such as part of speech while the values (i.e., *a* to *n*) represent specific parts of speech (e.g., noun, adjective, adverb, or verb, as per conventional grammatical descriptions) that have been acquired. For instance, adverbs will be subcategorized as ⟨A:h⟩ whereas most adjectives will be classified as either ⟨A:i⟩ or ⟨A:n⟩ as both clusters 9 and 14 mainly list adjectives.

From a more formal perspective, the feature can be regarded as a dimension with the specific classification representing a value within that dimension: Each requested language acquisition procedure introduces a major classification of linguistic knowledge such as part of speech or grammatical number into the grammar employed by the daughter language model. Conversely, the specific results of language acquisistion indicate which grammatical knowledge has been acquired with respect to this major category, for example, noun or second person. Then, each requested acquisition simulation either introduces a new type of linguistic knowledge to the grammar such as functional vs. content categories for a grammar that is already specifying part of speech and number or adds additional information with respect to previously acquired grammatical classifications. For instance, a new learning procedure could also model the acquisition of grammatical knowledge related to part of speech: In this case, the acquired categories should also be classified as values within this major classification, namely feature A. Conversely, if subsequent acquisition procedures are designed to acquire knowledge of another type of grammatical classification such as the distinction between function and content words, a different feature (e.g., B) should be used to specify the results in the daughter grammar. Then, the values for this feature — such as *a* and *b* representing function word vs. content word — correspond to the specific categories that could be acquired.

Crucially, in the MODOMA daughter language model formalism, a graph structure encoding a grammatical item can only be indicated for a single value for each feature or major linguistic category. Conversely, it can simultaneously be specified for multiple grammatical features.

For instance, the lemma for *onderwijs* ('education') can be specified as ⟨A:d⟩ (corresponding to ⟨PARTOFSPEECH:noun⟩) and ⟨B:b⟩ (e.g., indicating ⟨FUNCTIONORCONTENT:content⟩) simultaneously. To the contrary, it is impossible to classify this lemma as both ⟨A:a⟩, which corresponds to ⟨PARTOFSPEECH:verb⟩ and ⟨A:d⟩, which is associated with the grammatical category indication ⟨PARTOFSPEECH:verb⟩. Thus, this formalism enables the specification of a language model in which grammatical items are explicitly defined in terms of the acquired features while preventing contradictory classifications. This example illustrates the added value of using feature-value pairs rather than single labels to indicate acquired grammatical categories.

Moreover, by encoding the results of acquisition by feature-value pairs in the grammar, these properties can be used as restrictions on the combinatory properties of words and constructions during parsing and generating new utterances. Thus, the daughter agent has acquired categories it can use to generate grammatical utterances (according to the current grammar) and determine whether utterances generated by the mother language model are grammatical. In accordance with the interaction model, that is, acquisition is based on an ongoing interaction between the mother and daughter agents rather than a previously assembled corpus, the MODOMA system is designed in such a way that as soon as new grammatical knowledge has been acquired, it is employed to take part in the exchanging of sentences with the mother. It is a noteworthy characteristic of the MODOMA that unlike many existing language acquisition algorithms knowledge is used before the grammar has been fully acquired. As a result, the utterances produced by the daughter language model based on newly acquired grammatical items are instantly input to feedback by the mother language model, if requested by the users of the system. Thus, an assessment of the quality of the newly acquired grammatical knowledge can be carried out by the daughter language model.

7 Evaluating the Suggested Default Settings on Another Dataset

A second experiment was conducted using a test set of 10,000 sentences. The sentences were independently generated during a separate session by DELILAH employing the same settings as those determined by the training data. Thus, the proposed parameter settings were validated on an independent dataset. The structure of the results of hierarchical agglomerative cluster analysis on this test set containing 10,000 DELILAH utterances are presented by Figure 3 below. A full-size version of these results, which facilitates a more detailed inspection of individual data items and their relationships, is provided in Online Appendix 2 in the ancillary files. We will highlight some major and noteworthy groupings identified in this analysis. However, this account is not exhaustive. In this dendrogram, the first major split groups sentential adverbs to the left. This clustering of a specific subclass of adverbs is a non-trivial result. It is followed in the graph to the right by a group of words that take a complement clause. Most of them are infinitive or third person plural verbs such as *horen*, 'hear', and *zeggen*, 'say', except for the present participle *lezende* ('reading') and two nouns related to *vraag* ('question', *vraagje*, 'small question'). There are two subgroups that assemble nouns. Interestingly, the plural word forms connected to *vraag*, 'question', that is, *vragen*, a nominal meaning 'questions', which ambiguously also includes the verbal form *vragen*, 'question', and *vraagjes*, 'small questions', are grouped with the nouns here.

Moreover, there is a major split that primarily groups determiners, pronouns, and numerals. The first subcluster within this split also contains two verbs: *blijkt*, ('turns out') and *werkt*, ('works'). It is insightful that a subgroup assembles the personal pronouns *je* ('you' CLIT.INFORM), *jij* ('you', INFORM), and *u* ('you', FORMAL). Similarly, this cluster includes numerals and determiners such

as *acht* ('eight'), *twee* ('two'), and *sommige* ('some', PL) grouped together. Interestingly, the dendrogram in Figure 3 reveals that *massa's* ('masses') is grouped with the numerals and this correctly detects that the generator of DELILAH uses this word similarly to numerals. Thus, analysis by the MODOMA can increase knowledge of the results of processing by DELILAH. Within this cluster, numerals and determiners that are not numerals such as *die* ('that', MASC/FEM), *deze* ('this', MASC/FEM), *de* ('the', MASC/FEM/PL), *het* ('the', NEUT.SG), and *geen* ('no'), are placed together as well. Another insightful observation is that *elk* ('every', NEUT) and *ieder* ('every', NEUT) are grouped together in a subcluster as well as *elke* ('every', MASC/FEM) and *iedere* ('every', MASC/FEM).

In the dendrogram in Figure 3, there is a cluster that assembles functional categories. Moreover, the first subcluster in this group of functional categories includes only subjunctive conjunctions (e.g., *totdat*, 'until', *wanneer*, 'when', and *waarom*, 'why') while the next subgroup primarily lists prepositions such as *aan* ('to'), *in* ('in'), and *met* ('with'). Thus, these groupings provide valuable insights and are indicative of grammatical structures at multiple levels. A noteworthy major split consists of nominals, mainly adjectives such as *internationale* ('international'), *vervelende* ('annoying'), and *warme* ('warm') along with some nouns (e.g., *vrede*, 'peace', and *ethiek*, 'ethics') while another major cluster largely contains verbs. There are several subgroups within this major verb cluster: One subgroup consists especially of third person verbs such as *groeit* ('grows') and *functioneert* ('functions'). Two exceptions are the adjectives *kortdurende* ('short-term') and *sociaal-economische* ('socio-economic'). Moreover, a subgroup within the verb cluster assembles in particular various verbal forms such as third and plural person auxiliaries, for instance, *wordt*, 'becomes', and *worden*, 'become', plural forms (e.g., *huilden*, 'cried', PL), and past perfect participles (e.g., *gehouden*, 'held', and *gekomen*, 'come'). Notably, the infinitive complementizer *te* ('to') is also grouped with the verbal forms. Interestingly, within this group there are subclusters of auxiliaries: One such subgroup contains *worden*, 'become', *hebben*, 'have', and *zijn*, 'are', 'be', which is synonymous with the word for 'his', while another cluster includes *heeft*, 'has', and *waren*, 'were'. Within the latter, *is*, 'is', and *was*, 'was', are grouped together with a small distance to *waren*, 'were' in the dendrogram.

Similarly to the results of the hierarchical agglomerative clustering analysis of the training set, the test set containing 10,000 sentences also reveals a major cluster that corresponds less clearly to grammatical categories traditionally used in grammar descriptions. Nonetheless, several subgroups within this cluster primarily contain words that align with categories described in conventional grammar such as the groupings of *daar* ('there') and *hier* ('here'), the prepositions *over* ('about') and *uit* ('from'), and the possessives *mijn* ('my') and *uw* ('your', FORMAL). Another interesting observation is that there is a main split that largely contains infinite verb forms in an ordered manner. For instance, a subgroup within this cluster specifically includes perfect past participles (e.g., *gedaan*, 'done', and *geboden*, 'offered'). Other clusters in this major split predominantly group verbal forms as well such as *zittende* ('sitting'), *dienen* ('serve'), and *zitten* ('sit') while the following main split consists of third person plural or infinitive verbs including *beseffen* ('realize'), *hangen* ('hang'), and *willen* ('want'). Finally, a major split should be noted, which again encompasses nominals: These are mostly adjectives such as *zware* ('heavy'), *onhandige* ('clumsy'), and *algemene* ('general') as well as some nouns (e.g., *aanbod*, 'offer', *verwachtinkje*, 'small expectation', and *koffie*, 'coffee').

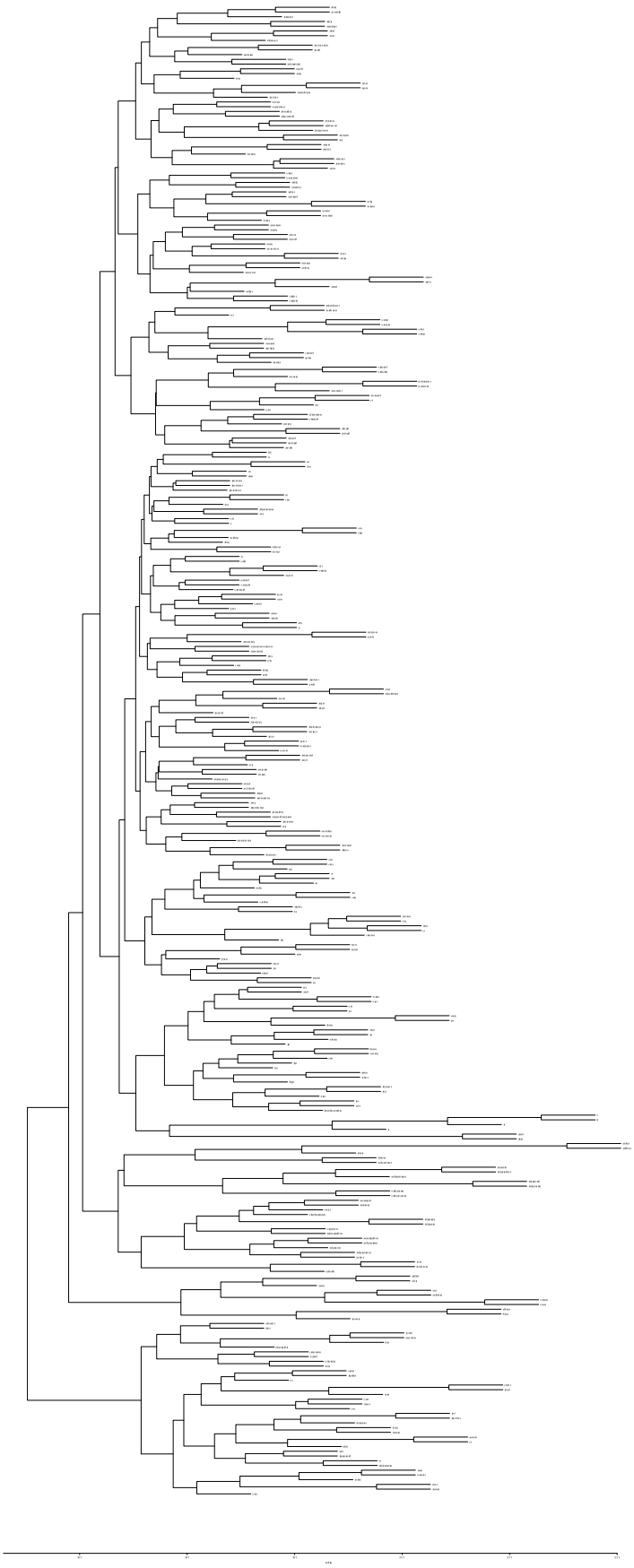

Figure 3: Structure of clustering based on 10,000 DELILAH exemplars for the data from session 160158110733 8. For a full-size dendrogram presenting all data in a format that allows for detailed inspection of individual items, consult Online Appendix 2 provided in the ancillary files.

Taking the results of the intermediate analysis based on hierarchical agglomerative clustering analysis, 14 grammatical categories have been requested employing the parameter settings determined by the first session. Table 5 presents the resulting categories. The translations and grammatical details with respect to these words have only been provided to facilitate the interpretation of the results but were not available to the unsupervised language processing procedures by the MODOMA. Interestingly, the second category in Table 5 contains functional categories, which consist of prepositions (e.g., *aan*, 'to', and *voor*, 'for'), subjunctive complementizers (e.g., *dat*, 'that', and *omdat*, 'because'), and the auxiliary *hadden* ('had', PL). Similarly, with the exception of the verb *werkt* ('works') category 6 corresponds to determiners. Categories 3, 7 and 9 primarily include nominals (i.e., adjectives as well as nouns) while categories 4, 5 and 11 only list nouns. Group 14 consists exclusively of verbs and cluster 13 contains mostly verbs. Additionally, category 8 includes mainly past participles along with some prepositions. Category 12 lists words that take a sentential complement, namely verbs and words connected to *vraag* ('question'). Finally, category 10 contains sentential adverbs. Conversely, the first group does not seem to correspond clearly to a category described by conventional grammars as it contains for instance adverbs such as *daar* ('there') and *dan* ('then', also meaning 'than'), adjectives (e.g., *binnenlands*, 'domestic', and *sociaal-economische*, 'socio-economic'), various verbal forms (e.g., *functioneert*, 'functions', and *zijn*, 'are', 'be', which is synonymous with the word for 'his'), possessives such as *mijn* ('my'), and prepositions (e.g., *over*, 'about', and *uit*, 'from').

#	FEATURE	VALUE	Words in the cluster
1.	A	a	*ben* ('am'), *binnenlands* ('domestic'), *daar* ('there'), *dan* ('then', 'than'), *directe* ('direct'), *er* ('there'), *ernstige* ('serious'), *functioneert* ('functions'), *gaan* ('go'), *gaat* ('goes'), *gehouden* ('held', PART), *gekomen* ('come', PART), *getoond* ('shown'), *groeit* ('grows'), *had* ('had', SG), *hangt* ('hangs'), *hebben* ('have'), *hebt* ('have', 2:SG), *heeft* ('has'), *hier* ('here'), *huilden* ('cried', PL), *hun* ('their'), *is* ('is'), *komt* ('comes'), *kortdurende* ('short-term'), *kwaadaardige* ('evil'), *meer* ('more'), *mijn* ('my'), *morgens* ('of the morgen', ARCHAIC GEN), *niet* ('not'), *over* ('about'), *'s* ('of the', ARCHAIC DET), *sociaal-economische* ('socio-economic'), *speelt* ('plays'), *te* ('at', 'to', COMP), *ter* ('at'), *terug* ('back'), *uit* ('from'), *uw* ('your', FORMAL), *verandert* ('changes'), *vereisende* ('demanding'), *verschijnt* ('appears'), *waren* ('were'), *was* ('was'), *wensende* ('wishing'), *werd* ('became', SG), *werden* ('became', PL), *worden* ('become'), *wordt* ('becomes'), *zijn* ('are', 'be', 'his')
2.	A	b	*aan* ('to', PREP), *als* ('if'), *alsof* ('as if'), *bij* ('at', 'near'), *dat* ('that'), *door* ('by'), *hadden* ('had', PL), *hoe* ('how'), *in* ('in'), *met* ('with'), *naar* ('to', PREP), *of* ('or', 'whether'), *omdat* ('because'), *op* ('on'), *tegen* ('against'), *terwijl* ('while'), *tot* ('until'), *totdat* ('until'), *van* ('of'), *vergeten* ('forgotten'), *voor* ('for'), *voordat* ('before'), *waar* ('where'), *waarom* ('why'), *wanneer* ('when'), *zoals* ('as'), *zodat* ('so that')

3.	A	c	*aanbod* (‘offer’), *actieve* (‘active’), *dramatische* (‘dramatic’), *gezamenlijke* (‘common’), *langdurige* (‘long-lasting’), *luie* (‘lazy’), *medische* (‘medical’), *museale* (‘museological’), *nationale* (‘national’), *omvangrijke* (‘extensive’), *onderwijs* (‘education’), *schone* (‘clean’), *verwachtinkje* (‘small expectation’), *zichtbare* (‘visible’), *zware* (‘heavy’)
4.	A	d	*aansporing* (‘incitement’), *beloftetjes* (‘little promises’), *besluitjes* (‘small decisions’), *eisjes* (‘small demands’), *initiatiefjes* (‘little initiatives’), *kans* (‘opportunity’), *kansen* (‘opportunities’), *mededelingen* (‘announcements’), *mededelingetjes* (‘little announcements’), *mogelijkheden* (‘possibilities’), *mogelijkheidjes* (‘small possibilities’), *opdracht* (‘assignment’), *opdrachten* (‘assignments’), *verwachtinkjes* (‘small expectations’), *voorstellen* (‘proposals’), *voorsteltjes* (‘small proposals’)
5.	A	e	*aansporingen* (‘incitements’), *feitjes* (‘small facts’), *kansjes* (‘small opportunities’), *vraagjes* (‘small questions’), *vragen* (‘questions’, N, ‘question’, V)
6.	A	f	*acht* (‘eight’), *alle* (‘all’), *blijkt* (‘turns out’), *de* (‘the’, MASC/FEM/PL), *deze* (‘this’, MASC/FEM), *die* (‘that’, MASC/FEM), *dit* (‘this’, NEUT), *drie* (‘three’), *een* (‘a’, ‘one’), *elk* (‘every’, NEUT), *elke* (‘every’, MASC/FEM), *enkele* (‘a few’), *geen* (‘no’), *het* (‘the’, NEUT:SG), *ieder* (‘every’, NEUT), *iedere* (‘every’, MASC/FEM), *ik* (‘I’), *je* (‘you’, INFORM), *jij* (‘you’, INFORM), *massa’s* (‘masses’), *menig* (‘many’, SG), *menige* (‘many’, PL), *negenennegentig* (‘ninetynine’), *sommige* (‘some’, PL), *twaalf* (‘twelve’), *twee* (‘two’), *u* (‘you’, FORMAL), *veel* (‘a lot of’), *vier* (‘four’), *vijf* (‘five’), *vijftig* (‘fifty’), *weinig* (‘a few’), *werkt* (‘works’)
7.	A	g	*adequate* (‘adequate’), *algemene* (‘general’), *arme* (‘poor’), *belangrijkere* (‘more important’), *dankbare* (‘grateful’), *etnische* (‘ethnic’), *goede* (‘good’), *grijze* (‘grey’), *haar* (‘her’), *halve* (‘half’), *initiatieven* (‘initiatives’), *koffie* (‘coffee’), *lange* (‘lang’), *multi-etnische* (‘multi-ethnic’), *nadrukkelijke* (‘explicit’), *onhandige* (‘clumsy’), *slaap* (‘sleep’), *spelen* (‘play’), *vitale* (‘vital’), *WAO* (‘disability law’)
8.	A	h	*af* (‘off’, ‘out’), *begrepen* (‘understood’, PART), *betreurd* (‘regretted’, PART), *bevorderd* (‘promoted’, PART), *gebleken* (‘seemed’, PART), *geboden* (‘offered’, PART), *gebouwd* (‘built’, PART), *gedaan* (‘done’, PART), *gehoord* (‘heard’, PART), *gekund* (‘able’, PAST PART of *kunnen*, ‘can’), *gemoeten* (‘had to’, PART), *gezegd* (‘said’, PART), *onderstreept* (‘underlines’, ‘underlined’, PART), *ontkend* (‘denied’, PART), *toe* (‘to’, ‘at’), *verantwoord* (‘responsible’), *voorkomen* (‘prevent’, ‘prevented’, PART), *voort* (‘forth’)

9.	A	i	*afschuwelijke* (‘horrible’), *agressieve* (‘aggressive’), *arbeidsongeschikte* (‘disabled’), *binnenlandse* (‘domestic’), *boze* (‘angry’), *duurzame* (‘durable’), *ethiek* (‘ethics’), *ethische* (‘ethical’), *functionele* (‘functional’), *gezonde* (‘healthy’), *grove* (‘gross’), *internationale* (‘international’), *langzame* (‘slow’), *last* (‘burden’), *lekkere* (‘tasty’), *minderjarige* (‘underage’), *Nederlandse* (‘Dutch’), *ongekende* (‘unprecedented’), *onmisbare* (‘indispensible’), *openbare* (‘public’), *preventieve* (‘preventive’), *rare* (‘strange’), *relaxte* (‘relaxt’), *rustige* (‘quiet’), *second* (‘second’)[5], *snelle* (‘quick’), *stapsgewijze* (‘gradual’), *uitwerking* (‘effect’), *veilige* (‘safe’), *verkeerde* (‘wrong’), *verre* (‘remote’), *vervelende* (‘annoying’), *vrede* (‘peace’), *warme* (‘warm’), *zelfde* (‘same’), *zieke* (‘ill’)
10.	A	j	*al* (‘already’), *altijd* (‘always’), *bijna* (‘almost’), *binnenkort* (‘soon’), *daarna* (‘thereafter’), *daarom* (‘therefore’), *echter* (‘however’), *eerst* (‘at first’), *eigenlijk* (‘actually’), *even* (‘for a bit’), *gaarne* (‘with pleasure’), *gedeeltelijk* (‘partly’), *ginds* (‘over there’), *inderdaad* (‘indeed’), *ineens* (‘suddenly’), *meteen* (‘at once’), *misschien* (‘maybe’), *momenteel* (‘at the moment’), *natuurlijk* (‘of course’), *nog* (‘yet’), *nooit* (‘never’), *nu* (‘now’), *ook* (‘also’), *opnieuw* (‘again’), *soms* (‘sometimes’), *tegelijkertijd* (‘at the same time’), *tevens* (‘also’), *thans* (‘at present’), *toch* (‘however’), *toen* (‘then’, ‘when’), *verder* (‘further’), *vooral* (‘especially’), *voorheen* (‘previously’), *waarschijnlijk* (‘probably’), *weer* (‘again’), *wel* (‘well’), *zo* (‘so’, ‘immediately’)
11.	A	k	*amputatietjes* (‘small amputations’), *behandelingen* (‘treatments’), *beschrijvingen* (‘descriptions’), *bestralingen* (‘irradiations’), *bestralinkjes* (‘small irradiations’), *operaties* (‘operations’, ‘surgeries’), *operatietjes* (‘small operations’, ‘small surgeries’)
12.	A	l	*begrijpen* (‘understand’), *horen* (‘hear’), *lezen* (‘read’), *lezende* (‘reading’), *merken* (‘notice’), *vraag* (‘question’), *vraagje* (‘small question’), *weten* (‘know’), *zeggen* (‘say’), *zien* (‘see’)
13.	A	m	*behoren* (‘belong’), *beseft* (‘realizes’), *blijkende* (‘turning out’), *dienen* (‘serve’), *durven* (‘dare’), *gewenst* (‘desired’, PART), *om* (‘to’), *proberen* (‘try’), *verlangen* (‘desire’), *vermogens* (‘abilities’, ‘assets’), *vermogentjes* (‘small abilities’, ‘small assets’), *zitten* (‘sit’), *zittende* (‘sitting’)

[5]This is an English loanword, which was used by DELILAH as part of the collocation *second opinion*.

14.	A	n	*beseffen* (‘realize’), *bidden* (‘pray’), *blijken* (‘turn out’), *concurreren* (‘compete’), *denken* (‘think’), *drentelen* (‘saunter’), *gebeuren* (‘happen’), *groeien* (‘grow’), *hangen* (‘hang’), *huilen* (‘cry’), *kleven* (‘stick’), *komen* (‘come’), *liggen* (‘lie’), *lijken* (‘seem’), *ontkennen* (‘deny’), *praten* (‘talk’), *raden* (‘guess’), *slapen* (‘sleep’), *snijden* (‘cut’), *staan* (‘stand’), *stijgen* (‘rise’), *veranderen* (‘change’), *verbeteren* (‘improve’), *wensen* (‘wish’), *werken* (‘work’), *willen* (‘want’), *winkelen* (‘shop’), *zwemmen* (‘swim’)

Table 5: Groupings resulting from the cluster analysis of the data from session 1601581107338

Many categories in Table 5 appear to correspond to the results of the previous experiment, which involved performing the same analysis with the same parameter settings on 10,000 training sentences generated by DELILAH as well. Not only are clusters corresponding to part of speech categories typically used in conventional grammar models such as noun, verb, and sentential adverb identified, but other grammatical categories are also distinguished for both datasets. For instance, the second group identified in the previous experiment corresponds to functional categories such as prepositions, complementizers, and a form of the auxiliary *hebben* (‘have’). Similarly, for both the training and test datasets categories containing verbs and the word *vraag* (‘question’), which take a sentential complement, as well as groups seemingly not corresponding to a category employed by more traditional grammars are discovered.

To quantitatively assess the correspondence between the categories resulting from the training and test sessions, we have created the crosstabulation presented in Table 6, which displays the number of overlapping lexical items for each possible combination of training and test categories. Subsequently, we have used Fisher’s exact test with 500,000 Monte Carlo simulations to determine whether the association between the training and test clusters is significant at $p < 0.05$. As this analysis resulted in p = 2×10^{-6}, the null hypothesis of no association between the training and test categories is rejected. To further assess which training and test categories contribute most to this association, pairwise Fisher’s exact tests with Bonferroni-correction were carried out using R and the RVAideMemoire package (Hervé, 2022) for each two-by-two subtable that is entailed by Table 6, see the study by MacDonald and Gardner (2000) for a similar approach employing pairwise chi-square tests. These post-hoc tests indicated that in particular the subtables containing acquired training and test categories with overlapping lexical items contributed to this significant result. For example, Fisher’s exact tests with Bonferroni-correction performed on the two-by-two clusters for training categories ⟨A:b⟩ and ⟨A:h⟩ and test categories ⟨A:b⟩ and ⟨A:j⟩ as well as on training categories ⟨A:e⟩ and ⟨A:j⟩ and test categories ⟨A:f⟩ and ⟨A:l⟩ resulted in significant p-values of respectively p = 1.328×10^{-11} and p = 8.667×10^{-5}. This analysis is consistent with the observation that categories acquired during both experiments appear to correspond to each other.

8 Conclusions and Suggestions for Future Research

This article presents experiments conducted to evaluate the MODOMA system. The MODOMA serves as a multi-agent computational laboratory environment for language acquisition simulations

	Training													
Test	A:a	A:b	A:c	A:d	A:e	A:f	A:g	A:h	A:i	A:j	A:k	A:l	A:m	A:n
A:a	3	12	0	0	1	8	5	0	2	0	0	1	1	0
A:b	0	23	0	0	0	2	1	0	0	0	0	0	0	0
A:c	1	0	0	2	0	0	0	0	4	0	0	0	0	1
A:d	0	0	1	3	0	0	0	0	0	0	4	0	0	0
A:e	0	0	1	0	0	0	0	0	0	2	1	0	0	0
A:f	0	0	0	0	33	0	0	0	0	0	0	0	0	0
A:g	1	0	0	0	0	1	0	0	2	0	0	0	0	0
A:h	0	0	0	0	0	0	7	0	0	0	0	0	7	0
A:i	0	0	0	1	0	1	1	0	7	0	0	0	0	2
A:j	0	0	0	0	0	3	1	30	0	0	0	1	0	0
A:k	0	0	0	0	0	0	0	0	0	0	3	0	0	0
A:l	0	0	0	0	0	0	0	0	0	8	0	0	1	0
A:m	0	0	0	0	0	2	1	0	0	0	1	7	0	0
A:n	9	0	0	0	0	1	0	0	0	0	0	11	0	0

Table 6: Crosstabulation of the overlap between the acquired categories during the training and test experiments

enabling controlled experimentation and analysis. It employs two language models: a mother agent and a daughter agent. Both language models utilize comprehensive representations of linguistic knowledge for example including grammatical categories as well as the phonological and semantic forms of lexical items. The adult agent is based on DELILAH. This is a language model, which has been independently developed by Cremers and Hijzelendoorn (1995/2025) and provides a generator and parser executing a grammar model of Dutch (cf. Cremers et al., 2014, for a detailed discussion of this language model). Conversely, the daughter agent is a novel language model, which has been specifically developed for the MODOMA. It employs a language acquisition device to acquire knowledge of the target language. The result is used to define a grammar model of the linguistic knowledge of the target language, which is explicitly encoded through graph structures. These graphs employ feature-value pairs to represent the acquired linguistic categories. Moreover, the daughter language model includes a generator and a parser, which can execute her currently acquired grammar. Thus, language acquisition results in a productive language model. Crucially, as the MODOMA provides a laboratory environment for language acquisition experiments, all factors involved in language acquisition such as the adult agent, the language acquiring agent, and the exchange of utterances are integral components in a single system. Accordingly, all aspects of the system can be configured by the user through parameter settings while all results are logged and can be retrieved. This design opens up new possibilities for further research into language acquisition. Building on the work by Shakouri et al. (2025), the experiments presented in this article serve to both evaluate and illustrate the potential of this design.

In comparison with large language models, this line of research provides an interesting additional perspective of language modelling. A major advantage of a computational language acquisition laboratory along the lines of the MODOMA is that it offers inherently transparent artificial intelligence agents: The explicit representations of grammar entail that all knowledge employed by the system to process language can be consulted. Moreover, as all components involved in language

acquisition including the adult and child agents are part of the simulation, the system implements a fully accountable model of first language acquisition. This creates the possibility of conducting experiments that would be impossible to perform using human subjects such as comparing diverse experimental conditions. Given the growing use of language models, studying the extent to which machine-generated language resembles human language patterns is becoming increasingly significant. The MODOMA explores this issue, which is exemplified by this study. Finally, a notable design aspect is that the daughter agent only employs unsupervised techniques to construct a model of the target language. This resembles human children acquiring their mother language, who do not receive direct information on the labels employed by the adults. Thus, this system expands the perspective on modelling language acquisition.

This article discusses the results of two experiments aimed at the acquisition of grammatical categories for instance consisting of nouns, verbs, or determiners employing cluster analysis. The first experiment was conducted to determine the parameter settings while the second was used to validate these settings. The specification of discrete graph-based representations through unsupervised methods presents significant computational challenges. Therefore, a hybrid approach was employed in both experiments involving statistical as well as rule-based techniques. The statistical technique selected for this analysis was hierarchical agglomerative clustering, which was applied to a dataset consisting of 10,000 sentences generated by mother language model, see Figure 2 and Online Appendix 1 in the ancillary files for the intermediate results of the first experiment. This analysis resulted in a continuous representation of the grammatical relationships between the words. However, the daughter language model is designed to explicitly represent discrete grammatical categories. Therefore, based on the results of the training experiment, 14 categories were defined grouping the most similar items. Using these selected parameter settings, the first experiment demonstrated the feasibility of modelling the acquisition of discrete grammatical categories. These categories were represented using feature-value pairs specifying graph structures and the result was used to update the grammar model employed by the daughter language model. Therefore, this analysis enabled the system to acquire an abstract discrete grammar model of the target language. In particular, similarly to how grammatical categories are employed by more conventional grammar models, these updated structures indicate restrictions on the grammaticality of combinations of linguistic structures. Thus, these categories can be used productively by a language model employing a parser and/or a generator to generate grammatical structures and evaluate the grammaticality of input utterances depending on the currently acquired grammar. Non-trivially, the resulting categories largely correspond to distinctions made by more traditional grammatical models constructed by linguists, as shown in Table 3.

Using the default settings from the initial experiment, a new experiment was conducted to analyze 10,000 sentence utterances generated by the mother language model. Hence, hierarchical agglomerative clustering should yield 14 discrete categories for this second experiment as well. The intermediate results of this second simulation are displayed in Figure 3 and Online Appendix 2 in the ancillary files while the resulting 14 categories are presented in Table 5. Similarly to the first experiment, these categories often correspond to those used in conventional grammar models. This further suggests that the methods and parameter settings used to identify the patterns and acquire the categories are non-trivial. In both experiments, the MODOMA successfully specifies a discrete language model that describes abstractions similar to those found in other accounts of language. Concomitantly, statistical analyses of the overlap in lexical items between the training and test categories (see Table 6) revealed a significant association between the categories acquired

as a result of both experiments at $p < 0.05$. Post-hoc testing further indicated that this association is primarily driven by the training and test categories with matching lexical items. Accordingly, these analyses substantiate the observation that the acquired categories identified in both experiments are related. Thus, using the same parameter settings, the results from the training experiment are successfully replicated in the context of the test data.

Summarizing, by taking a statistical analysis as an intermediate step the acquisition procedures resulted in an abstracting rule-based grammar that encodes grammatical knowledge by leveraging discrete categories represented through feature-value pairs. Moreover, as soon as this abstracting grammatical knowledge has been acquired, it is applied by the daughter language model in two ways:

(1) The newly acquired knowledge is recorded in the grammar by specifying the lemmas with phonological forms identified by the learning procedure as corresponding to an acquired category. These lemmas are subcategorized for these feature-value pairs, which impose constraints on the combinatory properties of the lexical entries during parsing and generation. Therefore, the results of acquisition determine the grammaticality of newly parsed and produced utterances. Crucially, this final step entails the acquisition of abstract rule-based grammatical knowledge, which enables the daughter language model to make grammaticality judgements.

(2) The daughter language immediately utilizes the updated lemmas to exchange utterances with the mother agent.

Crucially, this study demonstrated that the MODOMA model of first language acquisition enables the acquisition of discrete grammatical categories and this knowledge is respresented by the daughter language model in such a way that it can be used productively to generate and parse novel utterances. A major contribution of this study is that it validates the concept of computational modeling in first language acquisition through the use of multi-agent systems: The experiments presented in this article strongly suggest the feasibility of advancing research into the computational modelling of language acquisition using multi-agent systems and offer a basis for multiple future research possibilities in this area. For example, the results of acquisition can be evaluated based on feedback, if requested. This suggests an interesting direction for further exploration. Moreover, previously acquired grammatical knowledge can be used as input to subsequent acquisition procedures. This learning strategy has been called internal annotation in the context of the MODOMA, which is a type of self-supervised learning (e.g., Balestriero et al., 2023; Gui et al., 2024). The MODOMA laboratory environment is an all-encompassing system, which includes all aspects of language acquisition from the generation of the samples of the target language by the mother and statistical analyses to adding newly acquired knowledge to the grammar and executing successive and diverse language acquisition strategies. Therefore, the use of internal annotation is an effective strategy for a MODOMA: In addition to employing newly acquired grammatical rules to generate and parse sentences, acquired grammatical knowledge can also be used to enable further acquisition procedures. The findings presented in this article establish a foundation for further exploration of internal annotation, with a particular focus on a multi-agent model of first language acquisition. These aspects will be the focus of upcoming studies.

Acknowledgements

The authors wish to thank Maarten Hijzelendoorn for his help and advice. This publication is part of the MODOMA project, which was financed by the Dutch Research Council (NWO).

References

Alishahi, A., & Chrupała, G. (2009). Lexical category acquisition as an incremental process. In *Proceedings of the CogSci 2009 workshop on Psycho Computational Models of Human Language Acquisition.*

Alishahi, A., & Chrupała, G. (2012). Concurrent acquisition of word meaning and lexical categories. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 643–654). Association for Computational Linguistics (ACL). Retrieved from https://aclanthology.org/D12-1059/

Alishahi, A., & Stevenson, S. (2008). A computational model of early argument structure acquisition. *Cognitive Science*, *32*(5), 789–834. doi: https://doi.org/10.1080/03640210801929287

Baayen, R. H., Feldman, L. B., & Schreuder, R. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language*, *55*(2), 290–313. doi: https://doi.org/10.1016/j.jml.2006.03.008

Baldridge, J., & Kruijff, G.-J. M. (2003). Multi-modal combinatory categorial grammar. In A. Copestake & J. Hajič (Eds.), *10th Conference of the European Chapter of the Association for Computational Linguistics (EACL)* (pp. 211–218). Association for Computational Linguistics (ACL). Retrieved from https://aclanthology.org/E03-1036

Balestriero, R., Ibrahim, M., Sobal, V., Morcos, A., Shekhar, S., Goldstein, T., ... Goldblum, M. (2023). A cookbook of self-supervised learning. *Computing Research Repository (CoRR)*, *arXiv:cs.LG/2304.12210v2*. doi: https://doi.org/10.48550/arXiv.2304.12210

Beekhuizen, B., Bod, R., Fazly, A., Stevenson, S., & Verhagen, A. (2014). A usage-based model of early grammatical development. In V. Demberg & T. O'Donell (Eds.), *Proceedings of the Fifth Workshop on Cognitive Modeling and Computational Linguistics (CMCL)* (pp. 46–54). Association for Computational Linguistics (ACL). doi: https://doi.org/10.3115/v1/W14-2006

Bod, R., Scha, R., & Sima'an, K. (2003). *Data-oriented parsing*. Stanford, CA: CSLI.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.-F. Balcan, & H.-T. Lin (Eds.), *Advances in neural information processing systems 33 (NeurIPS 2020)* (Vol. 33, pp. 1877–1901). Curran Associates. Retrieved from https://papers.nips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

Chaabouni, R., Kharitonov, E., Bouchacourt, D., Dupoux, E., & Baroni, M. (2020). Compositionality and generalization in emergent languages. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 4427–4442). Association for Computational Linguistics (ACL). doi: https://doi.org/10.18653/v1/2020.acl-main.407

Chaabouni, R., Kharitonov, E., Dupoux, E., & Baroni, M. (2019). Anti-efficient encoding in emergent communication. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. B. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)* (Vol. 32, pp. 6261–6271). Curran Associates. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2019/file/31ca0ca71184bbdb3de7b20a51e88e90-Paper.pdf

Chaabouni, R., Kharitonov, E., Dupoux, E., & Baroni, M. (2021). Communicating artificial neural networks develop efficient color-naming systems. *Proceedings of the National Academy of*

Sciences (PNAS), *118*(12). doi: https://doi.org/10.1073/pnas.2016569118

Chaabouni, R., Strub, F., Altché, F., Tallec, C., Trassov, E., Davoodi, E., ... Piot, B. (2022). Emergent communication at scale. In *Proceedings of the Tenth International Conference on Learning Representations (ICLR 2022).* Retrieved from https://openreview.net/pdf?id=AUGBfDIV9rL

Chomsky, N. (1963). Formal properties of grammar. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. II, pp. 323–418). New York, NY: Wiley.

Conner, M., Gertner, Y., Fisher, C., & Roth, D. (2008). Baby SRL: Modeling early language acquisition. In A. Clark & K. Toutanova (Eds.), *Proceedings of the Twelfth Conference on Computational Natural Language Learning (CoNLL 2008)* (pp. 81–88). Coling 2008 Organizing Committee. Retrieved from https://aclanthology.org/W08-2111/

Conner, M., Gertner, Y., Fisher, C., & Roth, D. (2009). Minimally supervised model of early language acquisition. In S. Stevenson & X. Carreras (Eds.), *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009)* (pp. 84–92). Association for Computational Linguistics (ACL). Retrieved from https://aclanthology.org/W09-1112

Cremers, C. (2002). ('n) Betekenis berekend. *Nederlandse Taalkunde*, *7*(4), 375–395. Retrieved from https://journal-archive.aup.nl/nederlandse-taalkunde/taalk_2002_nr4.pdf

Cremers, C., & Hijzelendoorn, P. M. (1995/2025). *Delilah. Computer software.* Department of General Linguistics/LUCL, Leiden University.

Cremers, C., & Hijzelendoorn, P. M. (2025). *Delilah does Dutch. Online demo version.* Retrieved from https://delilah.universiteitleiden.nl/indexen.html (Accessed on 15 January 2025)

Cremers, C., Hijzelendoorn, P. M., & Reckman, H. G. B. (2014). *Meaning versus grammar*. Leiden: Leiden University Press. doi: https://doi.org/10.1515/9789400601833

Cremers, C., & Reckman, H. G. B. (2008). Exploiting logical forms. In S. Verberne, H. van Halteren, & P.-A. J. M. Coppen (Eds.), *Computational Linguistics in the Netherlands 2007 (CLIN 18)* (Vol. 11, pp. 5–20). LOT. Retrieved from https://www.clinjournal.org/CLIN_proceedings/XVIII/cremers.pdf

Daelemans, W., & van den Bosch, A. (2005). *Memory-based language processing*. Cambridge: Cambridge University Press. doi: https://doi.org/10.1017/CBO9780511486579

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *North American Chapter of the Association for Computational Linguistics: Human Language Technologies 2019 (NAACL-HLT)* (Vol. 1 (Long and Short Papers), pp. 4171–4186). Association for Computational Linguistics (ACL). doi: https://doi.org/10.18653/v1/N19-1423

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification* (2nd ed.). New York, NY: Wiley.

Griffith, T. L., & Kalish, M. L. (2007). Language evolution by iterated learning with Bayesian agents. *Cognitive Science*, *31*(3), 441–480. doi: https://doi.org/10.1080/15326900701326576

Gui, J., Chen, T., Zhang, J., Cao, Q., Sun, Z., Luo, H., & Tao, D. (2024). A survey on self-supervised learning: Algorithms, applications, and future trends. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–20. doi: https://doi.org/10.1109/TPAMI.2024.3415112

Hervé, M. (2022). *RVAideMemoire: Testing and plotting procedures for biostatistics (Version 0.9-81-2). R package.* doi: https://doi.org/10.32614/CRAN.package.RVAideMemoire

ter Hoeve, M., Kharitonov, E., Hupkes, D., & Dupoux, E. (2022). Towards interactive language modeling. *Computing Research Repository (CoRR)*, *arXiv:cs.CL/2112.11911v2*. doi: https://doi.org/10.48550/arXiv.2112.11911

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). ALBERT: A lite BERT for self-supervised learning of language representations. In *Eighth International Conference on Learning Representations (ICLR 2020).* Retrieved from `https://openreview.net/pdf?id=H1eA7AEtvS`

MacDonald, P. L., & Gardner, R. C. (2000). Type I error rate comparisons of post hoc procedures for "IxJ" chi-square tables. *Educational and Psychological Measurement*, *60*(5), 735–754. doi: http://doi.org/10.1177/00131640021970871

Mächler, M., Rousseeuw, P. J., Struyf, A., Hubert, M., & Hornik, K. (2019). *Cluster: Cluster analysis basics and extensions (Version 2.1.0). R package.* doi: https://doi.org/10.32614/CRAN.package.cluster

MacWhinney, B. (2014). *The CHILDES project: Tools for analyzing talk* (3rd ed.). New York, NY: Routledge. doi: https://doi.org/10.4324/9781315805641

Marr, D. (1982/2010). *Vision: A computational investigation into the human representation and processing of visual information*. Cambridge, MA: MIT Press. doi: https://doi.org/10.7551/mitpress/9780262514620.001.0001

Matusevych, Y., Alishahi, A., & Vogt, P. A. (2013). Automatic generation of naturalistic child-adult interaction data. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *35*, 2996–3001. Retrieved from `https://escholarship.org/uc/item/2t9414br`

McCoy, R. T., Frank, R., & Linzen, T. (2020). Does syntax need to grow on trees? Sources of hierarchical inductive bias in sequence-to-sequence networks. *Transactions of the Association for Computational Linguistics*, *8*, 125–140. doi: https://doi.org/10.1162/tacl_a_00304

Moortgat, M. J. (1997). Categorial type logics. In J. F. A. K. van Benthem & A. G. B. ter Meulen (Eds.), *Handbook of logic and language* (pp. 93–177). Amsterdam: Elsevier.

Orhan, A. E., Gupta, V. V., & Lake, B. M. (2020). Self-supervised learning through the eyes of a child. In H. Larochelle, M. Ranzato, R. Hadsell, M.-F. Balcan, & H.-T. Lin (Eds.), *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)* (Vol. 33, pp. 9960–9971). Curran Associates. Retrieved from `https://proceedings.neurips.cc/paper_files/paper/2020/file/7183145a2a3e0ce2b68cd3735186b1d5-Paper.pdf`

Pollard, C. J., & Sag, I. A. (1987). *Information-based syntax and semantics* (Vol. 1: Fundamentals). Stanford, CA: Center for the Study of Language and Information.

Pollard, C. J., & Sag, I. A. (1994). *Head-Driven Phrase Structure Grammar*. Stanford, CA and Chicago, IL: Center for the Study of Language and Information and University of Chicago Press.

R Core Team. (2020). *R: A language and environment for statistical computing (Version 3.6.3). Computer software.* Vienna: R Foundation for Statistical Computing. Retrieved from `http://www.R-project.org/`

Reckman, H. G. B. (2009). *Flat but not shallow: Towards flatter representations in deep semantic parsing for precise and feasible inferencing* (PhD thesis, LOT Dissertation Series 203, Leiden University). Utrecht: LOT. Retrieved from `https://www.lotpublications.nl/flat-but-not-shallow-flat-but-not-shallow-towards-flatter-representations-in-deep-semantic-parsing-for-precise-and-feasible-inferencing`

Rita, M., Chaabouni, R., & Dupoux, E. (2020). "LazImpa": Lazy and Impatient neural agents learn to communicate efficiently. In R. Fernández & T. Linzen (Eds.), *Proceedings of the 24th Conference on Computational Natural Language Learning (CoNLL 2020)* (pp. 335–343). Association for Computational Linguistics (ACL). doi: https://doi.org/10.18653/v1/2020.conll-1.26

Schütze, H. (1995). Distributional part-of-speech tagging. In S. P. Abney & E. W. Hinrichs (Eds.), *Proceedings of the Seventh Conference of the European Chapter of the Association for Computational Linguistics (EACL '95)* (pp. 141–148). San Francisco, CA: Morgan Kaufman. doi: https://doi.org/10.3115/976973.976994

Shakouri, D. P., Cremers, C., & Schiller, N. O. (2025). A knowledge-based language model: Deducing grammatical knowledge in a multi-agent language acquisition simulation. *Computational Linguistics in the Netherlands Journal*, *14*, 167–189. Retrieved from `https://www.clinjournal.org/clinj/article/view/193`

Steedman, M. (1996). *Surface structure and interpretation*. Cambridge, MA: MIT Press.

Steels, L. (2000). The emergence of grammar in communicating autonomous robotic agents. In W. Horn (Ed.), *Proceedings of the 14th European Conference on Artificial Intelligence 2000 (ECAI 2000)* (pp. 764–769). Amsterdam: IOS. Retrieved from `https://frontiersinai.com/ecai/ecai2000/pdf/p0764.pdf`

Steels, L. (2001). Language games for autonomous robots. *IEEE Intelligent Systems*, *16*(5), 16–22. doi: https://doi.org/10.1109/5254.956077

Steels, L. (2015). *The Talking Heads experiment: Origins of words and meanings*. Berlin: Language Science. Retrieved from `https://doi.org/10.17169/FUDOCS_document_000000022455`

Steels, L., & Beuls, K. (2017). How to explain the origins of complexity in language: A case study for agreement systems. In S. S. Mufwene, C. Coupé, & F. Pellegrino (Eds.), *Complexity in language: Developmental and evolutionary perspectives* (pp. 30–47). Cambridge: Cambridge University Press. doi: https://doi.org/10.1017/9781107294264.002

Steels, L., van Eecke, P., & Beuls, K. (2018). Usage-based learning of grammatical categories. In M. Atzmueller & W. Duivesteijn (Eds.), *Preproceedings of the 30th Benelux Conference on Artificial Intelligence (BNAIC 2018)* (pp. 253–264). 's-Hertogenbosch: Jheronimus Bosch Academy of Data Science (JADS).

Steels, L., & Kaplan, F. (2001). AIBO's first words: The social learning of language and meaning. *Evolution of Communication*, *4*(1), 3–32. doi: https://doi.org/10.1075/eoc.4.1.03ste

Steels, L., & Loetzch, M. (2012). The grounded naming game. In L. Steels (Ed.), *Experiments in cultural language evolution* (Vol. 3, pp. 41–60). Amsterdam: John Benjamins. doi: http://doi.org/10.1075%2Fais.3.04ste

van Trijp, R. (2016). *The evolution of case grammar*. Berlin: Language Science. doi: https://doi.org/10.17169/langsci.b52.182

Urbanek, S. (2009). How to talk to strangers: Ways to leverage connectivity between R, Java and Objective C. *Computational Statistics*, *24*(2), 303–311. doi: https://doi.org/10.1007/s00180-008-0132-x

Urbanek, S. (2012). *JavaGD: Java graphics device (Version 0.6-1). R package.* doi: https://doi.org/10.32614/CRAN.package.JavaGD

Urbanek, S. (2017). *RJava: Low-level R to Java interface (Version 0.9-9). R package.* doi: https://doi.org/10.32614/CRAN.package.rJava

Wickham, H. (2011). The split-apply-combine strategy for data analysis. *Journal of Statistical*

Software, *40*(1), 1–29. doi: https://doi.org/10.18637/jss.v040.i01

Wickham, H. (2020). *Plyr: Tools for splitting, applying and combining data (Version 1.8.6). R package.* doi: https://doi.org/10.32614/CRAN.package.plyr

Wickham, H., & Grolemund, G. (2016). *R for data science: Import, tidy, transform, visualize, and model data* (1st ed.). Sebastopol, CA: O'Reilly Media.

Wickham, H., Vaughan, D., & Girlich, M. (2020). *Tidyr: Tidy messy data (Version 1.1.1). R package.* doi: https://doi.org/10.32614/CRAN.package.tidyr

Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics* (pp. 189–196). Association for Computational Linguistics (ACL). doi: https://doi.org/10.3115/981658.981684

Appendix A. Glossary of Grammatical Terms and Abbreviations

Term / Abbreviation	Description
1	first person
2	second person
3	third person
ADJ	adjective
ARCHAIC	archaic
CLIT	clitic
COMP	complementizer
DET	determiner
FEM	feminine
FORMAL	formal
GEN	genitive
INF	infinitive
INFORM	informal
MASC	masculine
N	noun
NEUT	neuter
PART	participle
PAST	past
PL	plural
PREP	preposition
PRES	present
SG	singular
V	verb

Table A.1: Glossary of grammatical abbreviations