A knowledge-based language model: deducing grammatical knowledge in a multi-agent language acquisition simulation

Shakouri, D.P.; Cremers, C.L.J.M.; Schiller, N.O.

Computational Linguistics in the Netherlands Journal 14 (2025) 167-189 Submitted 12/2024; Published 07/2025

A Knowledge-Based Language Model: Deducing Grammatical Knowledge in a Multi-Agent Language Acquisition Simulation

David Ph. Shakouri[*,**] D.P.SHAKOURI@HUM.LEIDENUNIV.NL
Crit Cremers[*] C.L.J.M.CREMERS@HUM.LEIDENUNIV.NL
Niels O. Schiller[*,**,***] NIELS.SCHILLER@CITYU.EDU.HK

[*] *Leiden University Centre for Linguistics (LUCL), Leiden University, the Netherlands*

[**] *Leiden Institute for Brain and Cognition (LIBC), Leiden University, the Netherlands*

[***] *City University of Hong Kong (CityU), Hong Kong*

Abstract

This paper presents an initial study performed by the MODOMA system. The MODOMA is a computational multi-agent laboratory environment for unsupervised language acquisition experiments such that acquisition is based on the interaction between two language models, an adult and a child agent. Although this framework employs statistical as well as rule-based procedures, the result of language acquisition is a knowledge-based language model, which can be used to generate and parse new utterances of the target language. This system is fully parametrized and researchers can control all aspects of the experiments while the results of language acquisition, that is, the acquired grammatical knowledge, are explicitly represented and can be consulted. Thus, this system introduces novel possibilities for conducting computational language acquisition experiments. The experiments presented by this paper demonstrate that functional and content categories can be acquired and represented by the daughter agent based on training and test data containing different amounts of exemplars generated by the adult agent. Interestingly, similar patterns, which are well-established for human-generated data, are also found for these machine-generated data. As the procedures resulted in the successful acquisition of discrete grammatical categories by the child agent, these experiments substantiate the validity of the MODOMA approach to modelling language acquisition.

1. Introduction

This paper presents a study demonstrating the acquisition of grammatical knowledge by the MODOMA. The term MODOMA is an acronym for *moeder-dochter-machine* (Dutch for 'mother-daughter-machine'). This framework is aimed at providing a language acquisition laboratory, that is, a simulation environment for language acquisition experiments. On the one hand, all aspects of the system are parametrized so that the settings are user-controlled while on the other hand, all results and executed language acquisition procedures are logged and can be retrieved. The MODOMA integrates several characteristics that enable unique possibilities for language acquisition experiments. The MODOMA implements a multi-agent design modelling parent-child interaction such that both the parent and the child are language models in a single system. Both agents employ explicit representations of their grammatical knowledge, making the acquired knowledge and language processing retrievable. This explicit representation distinguishes the MODOMA system from other language learning systems, such as large language models (LLMs), which do not rely on such explicit knowledge structures. This combination of properties provides new opportunities for conducting computational experiments simulating first language acquisition. In a typical MODOMA experiment the input data to the language acquisition algorithm are interactively generated by the mother agent while the daughter agent constantly updates grammatical representations and takes

part in the interaction with the mother agent based on the currently acquired grammar. This design represents a novel approach as most often computational models of language acquisition are based on inputting corpora to the language acquisition procedures (e.g. Alishahi and Stevenson 2008, Alishahi and Chrupała 2012, Conner et al. 2009, Matusevych et al. 2013).

In this paper, rather than limiting the term "language model" to the typical usage in current NLP, where it usually refers to neural models trained on a next-word prediction objective (e.g., transformer-based models like GPT, cf. Brown et al. 2020, Radford et al. 2018, and BERT, cf. Devlin et al. 2019), we use it in a broader sense, encompassing rule-based approaches, statistical methods, and artificial neural networks as all of these approaches are employed to address common tasks in NLP. This distinction is important because the mother agent in our multi-agent system relies on explicit linguistic rules, while language acquiring agent incorporates statistical methods to infer a rule-based model for a linguistic phenomenon.

The study presented by this paper investigates whether the MODOMA daughter agent is able to acquire and represent functional and content categories by integrating an application of the Zipfian distribution in the MODOMA language acquisition system. Specifically, we examine whether a multi-agent language acquisition system can derive functional and content categories from machine-generated utterances and represent them through a rule-based grammar. To address this question, we assess whether these utterances exhibit statistical patterns similar to those in human language and determine the amount of input data necessary to detect such patterns. We then explore whether these statistical tendencies can be converted into grammatical rules. Finally, we validate the results by assessing the alignment of the inferred categories with the original grammar and evaluating the system's reliability in acquiring structured linguistic knowledge in an unsupervised manner.

The discussed language acquisition experiments are aimed at adding explicit grammatical knowledge to a lexicon. Subsequently, the daughter agent can use this acquired knowledge to take part in the conversation with the mother. To this end, the system employs statistical machine learning techniques, resulting in an explicitly specified grammatical representation. This contributes to existing systems based on statistical methods and neural networks by offering a more interpretable, rule-based approach to language modeling. By employing the principle of Zipf's law (Mandelbrot 1954, Zipf 1949), this study provides an essential robustness evaluation of the feasibility of implementing a multi-agent simulation of first language acquisition. Crucially, the purpose of this study is not to perform a binary classification of function and content words per se, but rather to employ this classification as a tool for assessing the validity of this novel approach to modelling first language acquisition. The advantage of using Zipf's law is that it offers a solid and reliable basis for evaluating the system's performance and components. This is a well-established practice. For example, Chaabouni et al. (2019) and Rita et al. (2020) employ Zipf's law to develop a multi-agent framework to model language evolution, see also the recent studies by Dębowski (2021), Takahashi and Tanaka-Ishii (2019), and Takamichi et al. (2024).

Importantly, the experiments in this study validate all components necessary to enable a multi-agent model of first language acquisition. Specifically, the acquisition of function and content words employs a hybrid language acquisition strategy, which is based on statistical tendencies that facilitate the acquisition of discrete grammatical rules. Therefore, these results support the modeling of more complex grammatical phenomena and demonstrate the feasibility of the MODOMA research agenda. A study building on these results is the work by Shakouri et al. (2025), which demonstrates how grammatical categories such as noun, verb, and preposition can be acquired using unsupervised techniques. In this paper two experiments illustrating this approach are presented: the first simulation discussed in Sections 6 and 7 has been applied to a training data set generated by the mother agent. These experiments were used to determine the parameter settings for the acquisition of this linguistic phenomenon. The second experiment taking a novel test data set generated by the adult language model as its input was used to validate this configuration, see Section 8. The results suggest that the MODOMA is able to acquire these grammatical categories and represent them by the grammar formalism of the daughter agent.

```
ID:A+B
|HEAD: |CONCEPT:the
|      |PHON:de
|      |QSEM:the
|      |SLF:the
|      |SYNSEM:CAT:det
|PHON:C
|PHONDATA:lijnop(de,A+B,[arg(right(1),0,D)],C)
|SLF: {{[E$F&(B+G)#H],[ ],[ ]},I@someˆFˆand(and(quant(F,the),
     and(H,entails1(F,decr))),and(I,entails(F,incr)))}
|SYNSEM: |CAT:np
|        |EMBEDCONCEPT:J
|        |EXAGGR:[number:plur,person:3]
|        |EXTTH:K [A+B,L]
|        |FREEARGS:M
|        |FUNCL:decr
|        |FUNCR:incr
|        |GENDER:N
|        |NOMTYPE:O
|        |NUMBER:plur
|        |PERSON:3
|        |QMODE:def
|        |REFMODE:P
|        |SEX:Q
|        |SUBCAT:noun
|        |VOICE:R
|TYPE:np\0∼[ ]/0∼[nˆ0#B+G]
|ARG: |ID:B+G
|     |HEAD:CONCEPT:J
|     |PHON:D
|     |SLF:E
|     |SYNSEM: |CAT:n
|     |        |EXTTH:K [B+G,L]
|     |        |FREEARGS:M
|     |        |GENDER:N
|     |        |NOMTYPE:O
|     |        |NUMBER:plur
|     |        |REFMODE:P
|     |        |SEX:Q
|     |        |VOICE:R
```

Figure 1: Example of the representation of the lexical item for *de* ('the.MASC/FEM') by Delilah

2. Description of the MODOMA

The MODOMA consists of two main components: (1) the mother agent and (2) the daughter agent. Delilah (Cremers and Hijzelendoorn 1995/2024a), the Leiden generator and parser of Dutch, is employed as the mother. Crucially, Delilah's grammatical knowledge is not based on a corpus. The parser as well as the generator of the mother language model execute the same grammar

consisting of predefined graph structures, which are comprised of attribute-value matrices. These graphs correspond to words and grammatical constructions and are explicitly specified for many grammatical properties such as the phonological form, concepts, logical meaning representation (cf. Cremers and Reckman 2008), grammatical number, grammatical person, and the syntactic category (e.g. determiner, noun, or verb). An example of a Delilah template is provided in Figure 1 illustrating Delilah's lexical and grammatical knowledge of the word *de* ('the.MASC/FEM')[1]. These graphs have been specified based on linguistic analyses of the Dutch lexicon and grammar, making Delilah's grammar closely tied to linguistic insights into Dutch. Thus, the linguistic knowledge and processing can be consulted. In this respect Delilah offers a complementary perspective, particularly when contrasted with the data-driven approach of large language models.

While Delilah's lexicon contains a large but finite list of words and constructions, these templates can be combined to form more complex utterances such as noun phrases and sentences. For example, the template for the article *de* ('the.MASC/FEM") can be used by Delilah to form noun phrases (e.g., *de vrouw*, 'the woman'). These templates are combined through the process of unification. This is a mathematical procedure, which verifies whether the templates contain conflicting features. If no conflicting values are found, it creates a template specified for all information contained in the original templates, see Schieber (2003) for a more detailed account of unification applied to grammatical representations. As shown in Figure 1, the complement position associated with the article *de* ('the.MASC/FEM') is constrained to only contain noun phrases, as indicated by the specification ARG|SYNSEM|CAT:n. Given the highly detailed nature of Delilah's representations, it lies beyond the scope of this paper to provide a full account of all features involved. However, Cremers et al. (2014, 259-265) provide an elaborate account of unification in the context of Delilah. In the application of unification to graph structures to produce and parse utterances, Delilah resembles head-driven phrase structure grammar (HPSG, Pollard and Sag 1986, 1994). Here the terms 'grammar' and 'lexicon' are used synonymously as there is no principled formal distinction between more concrete words and more abstract grammatical rules. Crucially, this language model executes a combinatory list grammar, which is related to combinatory categorial grammar (CCG, Cremers 2002, pp. 378-386, Cremers et al. 2014, pp. 115-137, cf. Baldridge and Kruijff 2003, Moortgat 1997, Steedman 1996). The characteristics of this language model are extensively discussed by Cremers et al. (2014) and Reckman (2009, pp. 25-75).[2] For the purposes of the MODOMA, Delilah, the mother agent, has been taken as is.

Conversely, the daughter language model has been constructed specifically as part of the MODOMA. Similar to Delilah, the daughter agent is a parser and generator but unlike the adult language model the grammatical representations are underspecified and become more specific as a result of the language acquisition procedures. A non-trivial resemblance between the adult and child systems is that both employ graph structures for representing grammatical knowledge and their parsers and generators execute their respective grammars by unifying lexical structures to output well-formed utterances and provide grammaticality judgements and parses with respect to input utterances based on the currently specified grammar. A crucial characteristic of the parser and generator of the child agent is that they can execute an incomplete grammar containing underspecified structures as similar to humans acquiring their first languages the daughter language model is required to take part in the conversation with the mother before acquisition has completed.

The daughter agent comprises three modules: (1) a parser and generator, the automaton, which execute (2) a grammar, which is specified as a result of acquisition by (3) a language acquisition device (cf. Shakouri et al. 2018). Crucially, the parser and generator are not modified by the language acquisition procedures. The result of language acquisition is explicit grammatical knowledge represented by newly acquired lexical entries and/or grammatical properties of lexical entries limiting their combinatory properties. A minimal example of a lexical entry is given by Figure 2.

1. Appendix A provides a glossary of grammatical abbreviations.
2. A demo version can be consulted at `https://delilah.universiteitleiden.nl/indexen.html` (Cremers and Hijzelendoorn 2024b), last accessed Oct. 1, 2024.

memory stack position:	6878
lexical entry number:	6879
session ID:	1731286472057
confidence lexical entry:	60
head directionality:	null
confidence head directionality:	0
terminal:	T
phonform:	*auto*
semform:	EKC
semform index:	null
GRAMMATICAL PROPERTIES:	[property type: A; property value: b; confidence property: 60]
HEAD:	[AUTO.6879]
ARGUMENT:	[null]

Figure 2: Example of the representation of an acquired lexical item for *auto* ('car') by the daughter agent

These data structures provide comprehensive representations of words and constructions. They have been specified for several grammatical properties, for example the phonological form (phonform), a representation of semantic content (semform), and an expandable list of other acquired grammatical properties. Grammatical properties are formalized by feature-value pairs such as [CONTENT OR FUNCTION WORD: content word] and [CONTENT OR FUNCTION WORD: function word] the acquisition of which is the aim of the study presented by this paper. As the learning procedures are unsupervised, the daughter language model has no information on the labels employed by the mother language model or grammar descriptions. Therefore, the acquired grammatical properties and the semform are indicated by unique alpha-numeric labels such as the property type 'A' and property value 'b' in Figure 2, see also Figure 3. These representations encode other grammatical properties not investigated by this paper as well. For example, the terminal feature differentiates between terminal words and more abstract constructions while the semform index enables representing anaphora such as reflexives. In addition, technical features are included such as the session id and the lexical entry number, which uniquely identify each rule. As a result of simulating language acquisition, the grammar can consist of increasingly complex structures as the HEAD and ARGUMENT can contain data structures identical to lexical items.

Since all linguistic knowledge is explicitly represented and can be consulted by researchers with respect to both agents, this type of approach provides an insightful addition to systems that learn language but do not employ explicit representations such as large language models. The MODOMA implements unsupervised learning as the input data to the language acquisition device are unannotated and this module has no access to the mother agent's internal processing and parses. Moreover, previously acquired grammatical knowledge by the daughter agent can be employed as input to the language acquisition procedures. This additional acquisition technique, which was introduced for the purposes of the MODOMA, has been named internal annotation. This is a form of self-supervised learning (cf. Balestriero et al. 2023, Brown et al. 2020, Devlin et al. 2019, Gui et al. 2024, Lan et al. 2020, Orhan et al. 2020, Yarowsky 1995).

3. Related work

As far as we know, the combination of characteristics and research possibilities entailed by the multi-agent MODOMA framework is a novel design. In particular, the resulting system can be used to cast language acquisition experiments based on natural language involving an adult language model and an acquiring language model. These agents take part in interactions, such that all aspects of the system, for example the parameter settings, input, executed language acquisition procedures, interaction, generated utterances, and acquired grammar, are user-controlled, reproducible, measurable, and verifiable. Ter Hoeve et al. (2022) present a road map of a teacher-student-loop to model language acquisition and implement two experiments corresponding to the first steps. The first experiment is aimed at modelling separate domains by 'two strictly separated vocabularies' of an artificial language while the second experiment addresses distinctive structures modeled by repetitions of tokens. As part of these experiments, the teacher selects a fixed number of training data from a set of pre-constructed sentences to provide them to multiple students, which consist of language models. After training, the student language models are examined with respect to a test set consisting of different sentences of the artificial language. In addition, a teacher language model is trained on another subset of sentences and subsequently assessed. The current implementation of this framework differs from the MODOMA most notably with respect to the training and test data: In the case of the work by ter Hoeve et al. (2022) the input contains pre-constructed utterances of an artificial language while for the MODOMA the data are representative of a natural language and produced online by Delilah as part of the interaction based on an explicitly defined grammar model of Dutch.

From an evolutionary perspective, research has been conducted on emergent communication utilizing multi-agent systems. An important example is the Talking heads framework. This project has been developed by Steels (2000, 2015) and provides a model of language evolution based on interactions including references to non-linguistic contexts between agents. These agents develop an artificial language as a result of language games (Steels 2001) such as the naming game aimed at defining names for objects in the surroundings (cf. Steels and Kaplan 2001 for studies on word naming, Steels and Loetzch 2012). Currently, their work also addresses the emergence and acquisition of syntactic phenomena (cf. Steels and Beuls 2017, Steels et al. 2018, van Trijp 2016). Relevant differences with the MODOMA project are that both agents engage in developing and learning language. Moreover, while the MODOMA daughter addresses the acquisition of (a fragment of a) natural language, the talking heads acquire an artificial language.

Chaabouni et al. (2020) study the emergence and beneficial properties of compositionality and generalization by two deep neural agents: A Receiver agent constructs a message m based on a received input i while a Sender agent is inputted with m and returns an output $\hat{\imath}$. Then, the game is successful if the input matches the output, that is, $i = \hat{\imath}$. Chaabouni et al. (2022) provide an architecture involving a Speaker, which receives an image and outputs a message, and a Listener, which aims at selecting the same image from a set of images based on the message, to study the effects of scaling up the 'data set, task complexity, and population size'. Although these studies involve two agents, unlike the MODOMA the language employed by these agents involves 'emergent codes' rather than (a fragment of) a natural language. Interestingly, Chaabouni et al. (2019) investigate whether two communicating agents can create an artificial language following Zipf's law, namely that more frequent words are likely to be shorter. To the contrary, they demonstrate that an anti-efficient code arises such that the most frequent words are correlated with the longest messages. Conversely, Rita et al. (2020) indicate an artificial language congruent with Zipf's law is learned for a LazImpa system involving a 'Lazy Speaker' and an 'Impatient Listener'. Griffith and Kalish (2007) model language evolution by (a population of) Bayesian agents executing iterative learning with respect to languages consisting of combinations of representations of meanings and utterances.

Alishahi carried out research simulating the acquisition of language by applying algorithms to corpora that have been human or artificially generated, see Alishahi and Chrupała (2009, 2012)

and Alishahi and Stevenson (2008) for studies modelling the acquisition of lexical categories, word meaning, and early argument structure. Matusevych et al. (2013) enhance artificially generated input corresponding to child-adult interactions to use them similarly to data such as the CHILDES database (MacWhinney 2014) but differently from the MODOMA framework as the algorithm that executes language acquisition, does not take part in the conversation. Beekhuizen et al. (2014) provide a different perspective by modelling the acquisition of grammatical constructions employing a single agent taking as its input generated representations of utterances and meanings. Furthermore, Conner et al. (2008, 2009)'s Baby SRL project models the acquisition of semantic role labelling by inputting annotated data from the CHILDES corpus or constructed sentences into their algorithm.

4. Modelling the acquisition of functional and content categories

The experiments presented in this paper assess whether a computational laboratory simulation of language learning could acquire function and content categories in an unsupervised fashion as a result of an interaction with Delilah, the language model providing samples of the adult language. Subsequently, the daughter language model can employ this newly acquired grammatical knowledge to take part in an interaction with the mother. On the one hand, the daughter agent can use the acquired grammar specified for these classifications to generate new sentences. These sentences can be presented to the mother as part of the conversation and the mother agent can return feedback (if requested by the parameter settings of the MODOMA) so that the daughter can assess the quality of the classifications. On the other hand, during the interaction the daughter uses this grammatical knowledge to parse novel utterances produced by the mother, for instance by classifying words presented by the mother with respect to this distinction.

Interestingly, it is a well-known finding in corpus linguistics that words corresponding to functional categories tend to occur with a high frequency whereas content words tend to have a low frequency of occurrence: many content words are used as hapax legomena, that is, only once, or a few times in a text, see Powers (1998) for an analysis of closed and open class words. This finding is congruent with Zipf's law, that is, the observation that the ranking of a word in a list of mostly used words is (approximately) proportional to the frequency of use for each word with respect to a particular factor (cf. Mandelbrot 1954, Zipf 1949). Moreover, a correlation between a distribution observed in languages and grammatical categories distinguished by linguists is suggested for language produced by humans. This paper assesses whether a similar correlation can be found for language produced by a machine.

An initial experiment implemented in the MODOMA laboratory simulation environment has been based on this correlation. This experiment provides a method to evaluate the MODOMA approach to language acquisition experiments employing the well-established principle of Zipf's law. Thus, this study contributes to a research development such that the MODOMA is gradually evolved by implementing increasingly complex experiments. As a result, experience with the MODOMA as a laboratory model for language acquisition experiments can be increased and the system can be tested and improved before implementing additional language acquisition techniques. Although Chaabouni et al. (2019) and Rita et al. (2020) focus on language emergence, there are some similarities with respect to the application of Zipf's law to increase understanding of language models in comparison with human-generated languages. Crucially, this procedure involves most of the components and procedures required by more intricate acquisition experiments. Accordingly, the component parts of the MODOMA system and their interaction, can be validated and improved to enable the implementation of further computational language acquisition experiments. If the acquisition succeeds, a distinction between functional and content categories is learned by the daughter agent based on samples of the target language presented by the mother agent during an interaction with the daughter. In this respect, it is significant that as soon as the daughter agent has acquired this grammatical distinction, the new grammatical knowledge is immediately used to update the daughter grammar and subsequently employed to generate novel utterances during the conversations with the mother

agent and parse novel mother sentences. Consequently, this study provides an important indication of whether the MODOMA laboratory approach to language acquisition can result in the acquisition of non-trivial grammatical knowledge that can be used productively during parsing and generation.

5. Methodology

Parameter	**Description**
Acquisition of functional categories	determines whether the system attempts to execute the acquisition of functional and content categories.
type of phrase generated by the mother agent	specifies whether Delilah should generate noun phrases (NPs) or full sentences when requested to produce utterances.
Minimum amount of data processed by the daughter agent	defines the minimum amount of data the system should process before starting the acquisition of functional and content categories.
Threshold for functional category acquisition	defines the threshold based on the type/token ratio to differentiate between functional and content categories in the utterances generated by the mother agent.

Table 1: Parameters for the functional and content category acquisition experiments

As the MODOMA implements a laboratory environment for language acquisition experiments, it is a fully parametrized system. This means that users can control the properties of each experiment by specifying predefined settings. Crucially, all aspects of language acquisition, that is, the adult, the daughter and the executed language procedures, are included in the system. Accordingly, the parameters can pertain to all these components. Table 1 provides an overview of the relevant parameters for the current experiments. The MODOMA is designed to execute several acquisition procedures (see also Shakouri et al. 2025, for an example of another acquisition procedure implemented using the MODOMA). The first parameter in Table 1, *Acquisition of functional categories*, specifies whether the acquisition of functional categories is enabled. In all experiments discussed in this paper, this is the case.

The remaining parameters define key experimental properties. The *type of phrase generated by the mother agent* parameter controls the structure of Delilah's utterances, determining whether sentences or noun phrases are generated. The training and test experiments presented in this paper systematically vary this parameter. Similarly, the *minimum amount of data processed by the daughter agent* before acquiring functional and content categories is varied across all experiments. These two parameters define the experimental setup. Specifically, in the training experiments, acquisition of functional and content categories is triggered after processing 1,000 noun phrases (NPs) and 1,000 sentences, followed by 10,000 noun phrases (NPs) and 10,000 sentences, respectively. This allows us to assess how many input exemplars generated by Delilah are required to detect this statistical tendency. Additionally, the *threshold for functional category acquisition* parameter specifies the type-token ratio used to convert distributional differences between function and content words into discrete linguistic knowledge. The training sessions discussed in this paper are used to determine appropriate values for these parameters, while the test experiments verify the selected settings using new datasets generated by Delilah.

Accordingly, Delilah was requested to generate eight sets of respectively 1,000 and 10,000 NPs and sentences. For each experiment, two of these sets were used to configure the acquisition procedures while two other data sets were employed to assess whether based on the same parameter settings similar results could be obtained. Initially, an experiment has been executed using the training sets of 1,000 NPs and sentences to determine whether diverging distributions between function and content words as established by the literature for human-generated corpora (cf. Powers 1998) could be obtained for utterances machine-generated by Delilah and whether this distinction is sufficiently

[A: a]	[CONTENT OR FUNCTION WORD: function word]
[A: b]	[CONTENT OR FUNCTION WORD: content word]

(a) Categories in the daughter grammar (b) Categories in conventional grammar

Figure 3: Representation of categories in the MODOMA versus conventional grammatical terminology

attested for the purposes of acquiring grammatical features. Subsequently, a follow-up experiment was performed with respect to the training and test data sets containing 10,000 NP and sentence exemplars produced by Delilah. If this distinction is detected by the language acquisition procedures in such a way that words can be reliably classified, the results are represented using graphs consisting of feature-value pairs specified by alpha-numeric labels, which should correspond to terminology more conventionally employed by linguists, compare Figure 3a with example labels applied by the MODOMA daughter agent and Figure 3b indicating more conventional labels.

6. Results

Table 2 presents lexical statistics of the training data containing 1,000 generated noun phrases and sentences by Delilah, including the total number of unique word types, the total number of word tokens, the number of hapax legomena (words that occur only once), and the number of word types that appear exactly twice. As an intermediate step during the acquisition procedures, for each experiment the word types have been sorted by frequency, see Table 3 and Table 4.[3] Interestingly, although this sample of 1,000 Delilah sentences contains more words than the collection of 1,000 NPs, the first unequivocal content word occurs at a higher rank in the list for 1,000 sentences, that is, *vergeten* ('forgotten') at rank 9. In accordance with the research hypothesis, the initial sets of 1,000 exemplars indicate a distinction between functional and content categories considering the most frequent words most often correspond to functional categories. However, this distinction is not clear enough for a system designed to acquire grammatical categories. Therefore, it is concluded that although data sets containing 1,000 exemplars generated by Delilah exhibit a distinction between function and content words similar to the one found for human language output, this difference is not substantiated sufficiently to enable the acquisition of grammatical features.

Accordingly, the follow-up experiment employed the larger data sets containing 10,000 NPs and sentences generated by Delilah. Table 5 summarizes the lexical statistics for these data sets. Similarly as for the previous experiment, the 35 most frequent words in both data sets are ranked in Table 6 and Table 7 respectively: As shown, these 35 highest-ranked word types are generally function words.

3. In these tables, f stands for frequency, that is, the number of tokens detected by the language acquisition procedures in the input data generated by the adult agent.

	NP training experiment	**Sentence training experiment**
Number of word types	1,319	1,728
Number of word tokens	3,027	5,874
Number of hapax legomena	776	859
Number of word types used twice	309	366

Table 2: Lexical statistics of the 1,000 generated NP and sentences training data

#	Word type	*f*
1.	*te* ('at', 'to', COMP)	92
2.	*dat* ('that')	62
3.	*niet* ('not')	60
4.	*om* ('to')	49
5.	*geen* ('no')	33
6.	*er* ('there')	32
7.	*de* ('the', MASC/FEM/PL)	25
8.	*elke* ('every', MASC/FEM)	25
9.	*een* ('a', 'one')	24
10.	*het* ('the', NEUT:SG)	24
11.	*op* ('on')	23
12.	*ieder* ('every', NEUT)	21
13.	*ik* ('I')	21
14.	*die* ('that', MASC/FEM)	20
15.	*elk* ('every', NEUT)	20
16.	*iedere* ('every', MASC/FEM)	20
17.	*veel* ('a lot of')	20
18.	*alle* ('all')	19
19.	*daar* ('there')	18
20.	*deze* ('this', MASC/FEM)	17
21.	*hier* ('here')	17
22.	*of* ('or', 'whether')	17
23.	*voor* ('for')	17
24.	*aan* ('to', PREP)	16
25.	*massa's* ('masses')	15
26.	*van* ('of')	14
27.	*vier* ('four')	14
28.	*weinig* ('a few')	14
29.	*dit* ('this', NEUT)	12
30.	*menig* ('many', SG)	12
31.	*negenennegentig* ('ninetynine')	12
32.	*blijkt* ('turns out')	11
33.	*naar* ('to', PREP, 'unpleasant')	11
34.	*twaalf* ('twelve')	11
35.	*vijf* ('five')	11

Table 3: 35 most frequent words for 1,000 generated NPs

#	Word type	*f*
1.	*dat* ('that')	197
2.	*te* ('at', 'to', COMP)	193
3.	*er* ('there')	119
4.	*niet* ('not')	91
5.	*ik* ('I')	75
6.	*hebben* ('have')	60
7.	*zijn* ('are', 'be', 'his')	50
8.	*om* ('to')	48
9.	*vergeten* ('forgotten')	41
10.	*alle* ('all')	38
11.	*deze* ('this', MASC/FEM)	38
12.	*elk* ('every', NEUT)	38
13.	*blijkt* ('turns out')	37
14.	*geen* ('no')	35
15.	*je* ('you', INFORM)	35
16.	*de* ('the', MASC/FEM/PL)	32
17.	*van* ('of')	32
18.	*worden* ('become')	32
19.	*door* ('by')	31
20.	*u* ('you', FORM)	31
21.	*jij* ('you', INFORM)	29
22.	*wordt* ('becomes')	29
23.	*hoe* ('how')	28
24.	*elke* ('every', MASC/FEM)	27
25.	*massa's* ('masses')	27
26.	*veel* ('a lot of')	27
27.	*werkt* ('works')	27
28.	*hier* ('here')	26
29.	*waarom* ('why')	25
30.	*ieder* ('every', NEUT)	24
31.	*weinig* ('a few')	23
32.	*aan* ('to', PREP)	22
33.	*iedere* ('every', MASC/FEM)	21
34.	*voor* ('for')	21
35.	*waar* ('where')	21

Table 4: 35 most frequent words for 1,000 generated sentences

In the sample of 10,000 NPs, the first unequivocal content word, *vergeten* ('forgotten'), appears at rank 54/55 sharing its position with the function word *bij* ('at', 'near'). After rank 55, the amount of content words increases. Similarly, in the experiment with 10,000 sentence training exemplars generated by Delilah, the most frequent content words are *vergeten* ('forgotten') at rank 11, *werkt* ('works') at position 27, and *morgens* (part of the archaic idiom *'s morgens* 'in the morning') at rank 66. After rank 66, the number of content words increases. Notably, Delilah uses the word *massa's* ('masses') as a numeral, which is accurately detected by these acquisition procedures.

	NP training experiment	Sentence training experiment
Number of word types	2,989	3,455
Number of word tokens	30,101	57,732
Number of hapax legomena	376	418
Number of word types used twice	396	355

Table 5: Lexical statistics of the generated 10,000 NP and sentences training data

#	Word type	*f*
1.	*te* ('at', 'to', COMP)	1018
2.	*dat* ('that')	658
3.	*niet* ('not')	650
4.	*om* ('to')	446
5.	*de* ('the', MASC/FEM/PL)	291
6.	*er* ('there')	271
7.	*geen* ('no')	270
8.	*elk* ('every', NEUT)	251
9.	*iedere* ('every', MASC/FEM)	247
10.	*of* ('or', 'whether')	244
11.	*elke* ('every', MASC/FEM)	229
12.	*ieder* ('every', NEUT)	223
13.	*alle* ('all')	215
14.	*een* ('a', 'one')	212
15.	*ik* ('I')	194
16.	*massa's* ('masses')	186
17.	*het* ('the', NEUT:SG)	180
18.	*op* ('on')	178
19.	*deze* ('this', MASC/FEM)	172
20.	*weinig* ('a few')	171
21.	*daar* ('there')	163
22.	*aan* ('to', PREP)	161
23.	*veel* ('a lot of')	155
24.	*hier* ('here')	146
25.	*voor* ('for')	144
26.	*dit* ('this', NEUT)	140
27.	*zijn* ('are', 'be', 'his')	130
28.	*door* ('by')	122
29.	*in* ('in')	120
30.	*menig* ('many', SG)	118
31.	*menige* ('many', PL)	111
32.	*je* ('you', INFORM)	110
33.	*die* ('that', MASC/FEM)	103
34.	*waar* ('where')	97
35.	*sommige* ('some', PL)	90

Table 6: 35 most frequent words for the 10,000 generated NPs training data

#	Word type	*f*
1.	*dat* ('that')	1999
2.	*te* ('at', 'to', COMP)	1779
3.	*er* ('there')	1370
4.	*niet* ('not')	969
5.	*ik* ('I')	729
6.	*hebben* ('have')	599
7.	*blijkt* ('turns out')	460
8.	*om* ('to')	434
9.	*alle* ('all')	409
10.	*zijn* ('are', 'be', 'his')	386
11.	*vergeten* ('forgotten')	385
12.	*elk* ('every', NEUT)	330
13.	*de* ('the', MASC/FEM/PL)	324
14.	*ieder* ('every', NEUT)	317
15.	*iedere* ('every', MASC/FEM)	317
16.	*door* ('by')	315
17.	*geen* ('no')	294
18.	*aan* ('to', PREP)	287
19.	*veel* ('a lot of')	281
20.	*elke* ('every', MASC/FEM)	280
21.	*hier* ('here')	280
22.	*u* ('you', FORM)	272
23.	*daar* ('there')	269
24.	*op* ('on')	268
25.	*weinig* ('a few')	268
26.	*massa's* ('masses')	262
27.	*werkt* ('works')	255
28.	*deze* ('this', MASC/FEM)	253
29.	*worden* ('become')	240
30.	*jij* ('you', INFORM)	238
31.	*je* ('you', INFORM)	226
32.	*wordt* ('becomes')	224
33.	*waar* ('where')	216
34.	*een* ('a', 'one')	213
35.	*is* ('is')	210

Table 7: 35 most frequent words for the 10,000 generated sentences training data

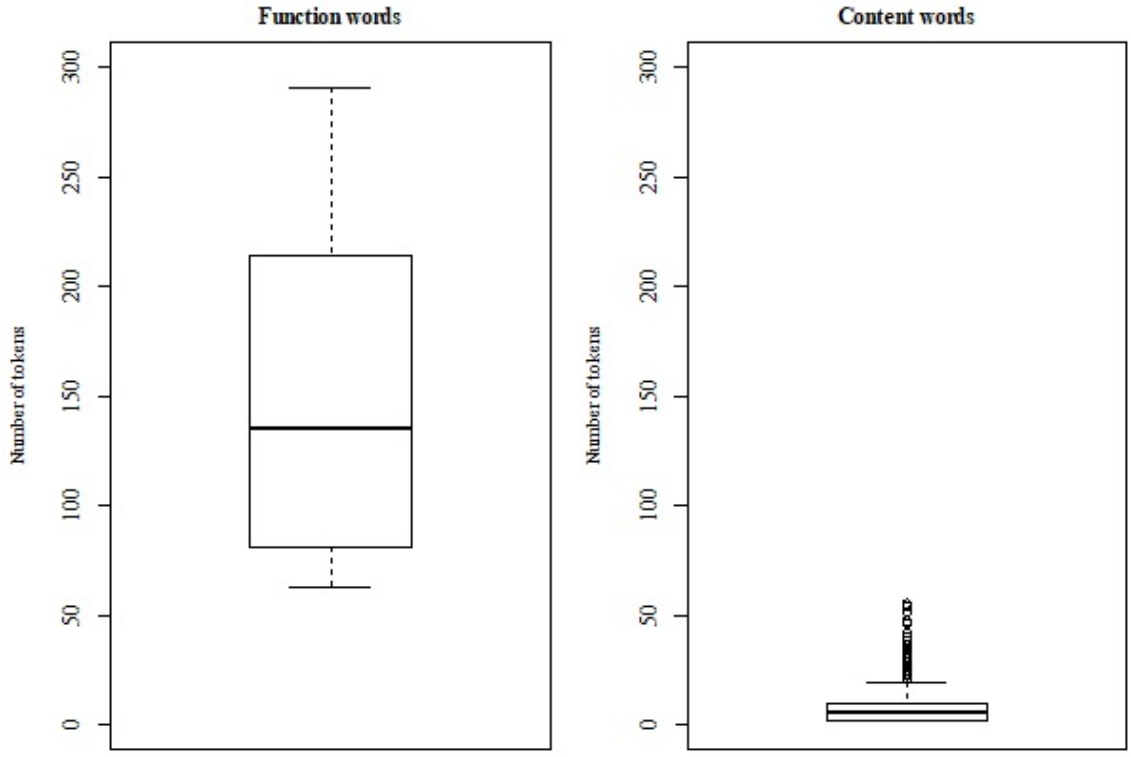

Figure 4: Boxplots depicting function and content words for the 10,000 NPs training data

7. Analyses

For data sets of 10,000 exemplars, a more straightforward difference between function and content categories can be manually detected such that words with a frequency of more than 100 tokens tend to be function words, that is, more than 100/57,732 or 1.7 per mil for the sample of Delilah sentences. Therefore, based on these experimentations the MODOMA parameters determining the acquisition of functional and content categories are set in order that word types that have a frequency of more than 2 per mil, are specified as functional categories while all other word types are considered content words. In addition, this acquisition procedure is executed only once after inputting at least 10,000 exemplars. These parameters are set conservatively, that is, in such a way that the chances of falsely classifying a word as functional category are smaller than assigning a specification as content word: As content word is the default category, each word should be considered a content word unless there is strong evidence for the opposite conclusion. The parameter values resulting from this initial experiment have been configured as the default settings for the MODOMA. However, these parameters can be adjusted depending on the experiment and taking into account previous experiments and linguistic theories.

Using the results of acquisition based on these settings, the boxplots in Figure 4 and Figure 5 have been made taking into account the frequency of both word types and these groupings suggest that for the NP as well as the sentence experiment the function words are distributed around other central values than the content words.[4] This provides further support for the research hypothesis and

4. The boxplots for the function words displayed outliers attested with a very higher frequency. Although these further confirm the patterns found that function words correspond to high rankings, these extreme values have been omitted as they would render the figures unreadable.

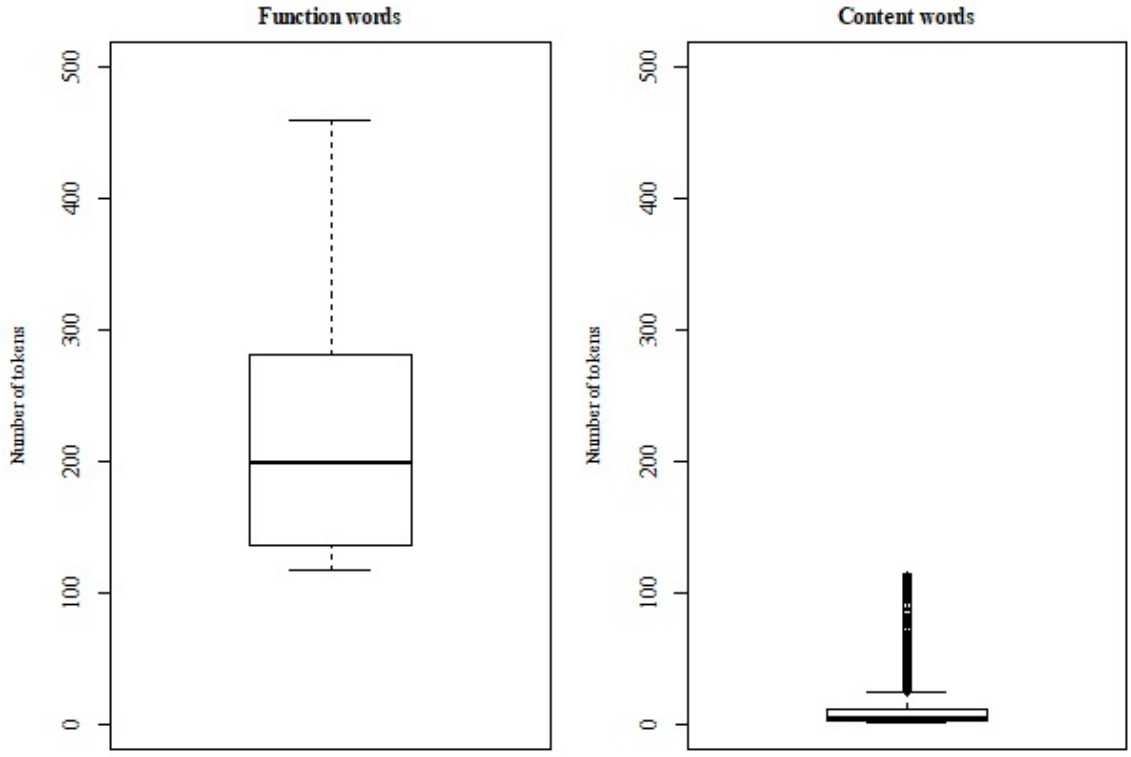

Figure 5: Boxplots depicting function and content words for the 10,000 sentences training data

the selected default parameter settings. Crucially, although the data are continuous, the acquired linguistic knowledge is discrete, that is, a word is either classified as a content or a functional category. These abstract categories explicitly specify the corresponding lexical items using a binary rule-based grammar formalism. This allows the daughter agent to apply the acquired feature-value pairs to generate and parse utterances based on unification during the conversation with the mother agent (cf. Pollard and Sag 1986, Pollard and Sag 1994 as well as Delilah for examples of unification-based language models, which do not implement language acquisition).

To quantitatively assess whether the acquired knowledge of the daughter agent matches the grammatical knowledge of the mother agent, a comparison has been made between the explicit grammatical specifications of both language models. Delilah does not directly specify for functional and content categories. Nonetheless, the mother language model employs categories such as transitive verb, count noun, or coordinating conjunction to process language, which correspond to function and content words. Therefore, the classifications in the grammar of the daughter resulting from the language acquisition simulations can be compared to the specifications in Delilah's core lexicon. This database contains the basic grammatical knowledge employed by the mother agent to generate the input exemplars to the language acquisition procedures. Fisher's exact tests have been executed to examine the relationship between a classification as function and content category by the mother and daughter language models. For all training experiments the results indicated there was a significant association between these variables, $p < 0.001$ (two-tailed). Further analyses revealed that in addition to matching classifications of both language models the mismatching classifications were predominantly function words according to the mother language model classified as content categories by the daughter agent.

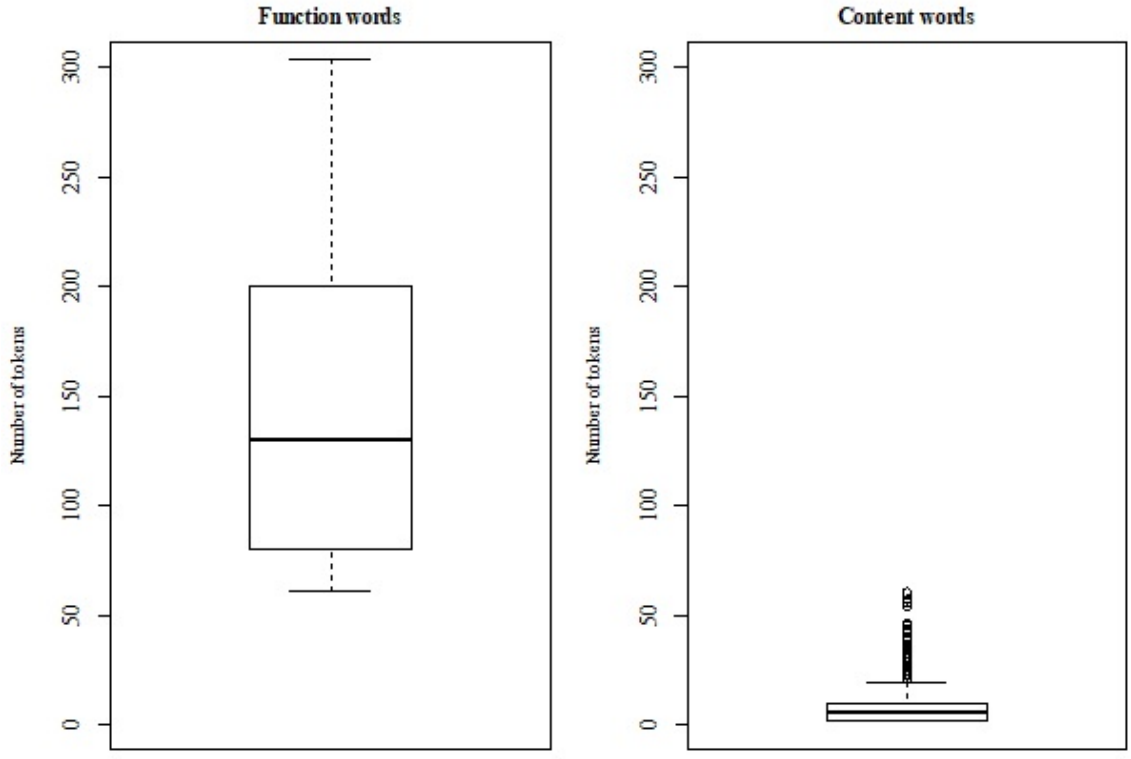

Figure 6: Boxplots depicting function and content words for the 10,000 NPs test data

8. Testing the suggested default settings using other data sets

To validate the parameter settings determined during the training sessions, the MODOMA has been requested to acquire function and content categories based on data sets not used to configure the parameter settings. To this end, two new data sets with respectively 10,000 NPs and sentences generated by Delilah have been created. Similarly as with respect to the training sets, the acquisition procedures for the test sets resulted in lists containing all word types in the data sets ranked in the order of the most frequent to the least frequent words as intermediate analyses. In accordance with the results of experimentation on the training data, the most frequent words correspond mostly to function words while the remainder of the words tend to be representative of content categories (according to more conventional grammar models). The 35 most frequent types occurring in these new data sets have been listed in Table 8 for the 10,000 NP sample and Table 9 for the exemplars representing 10,000 sentences.

Table 10 summarizes the distribution of word types, word tokens, and hapax legomena for the test sets containing 10,000 NPs and sentences. For the test set containing 10,000 NPs, the first content word in the corresponding list is *vragen* ('questions', N, 'question', V) at rank 56, which is attested with a frequency of 56. Therefore, the suggested default setting determined based on the experimentations with the training data, which predicts items occurring with a frequency greater than approximately 60 to be function words, is confirmed by the experiment assessing the NP test set. Similarly, with respect to the list constructed based on the 10,000 sentences test set, the first unequivocal content word is *weer* ('again', 'weather') at rank 50 with a frequency of 130 while the parameter settings configured by the training experiment predict the threshold to be at a frequency

#	Word type	*f*
1.	*te* ('at', 'to', COMP)	1025
2.	*niet* ('not')	633
3.	*dat* ('that')	631
4.	*om* ('to')	451
5.	*geen* ('no')	304
6.	*de* ('the', MASC/FEM/PL)	292
7.	*er* ('there')	261
8.	*ieder* ('every', NEUT)	253
9.	*elk* ('every', NEUT)	237
10.	*iedere* ('every', MASC/FEM)	231
11.	*elke* ('every', MASC/FEM)	224
12.	*of* ('or', 'whether')	215
13.	*een* ('a', 'one')	203
14.	*alle* ('all')	200
15.	*massa's* ('masses')	194
16.	*ik* ('I')	193
17.	*op* ('on')	178
18.	*aan* ('to', PREP)	173
19.	*deze* ('this', MASC/FEM)	172
20.	*het* ('the', NEUT:SG)	167
21.	*veel* ('a lot of')	162
22.	*weinig* ('a few')	160
23.	*voor* ('for')	145
24.	*hier* ('here')	143
25.	*die* ('that', MASC/FEM)	135
26.	*menige* ('many', PL)	131
27.	*door* ('by')	130
28.	*zijn* ('are', 'be', 'his')	130
29.	*dit* ('this', NEUT)	121
30.	*in* ('in')	120
31.	*daar* ('there')	114
32.	*menig* ('many', SG)	114
33.	*je* ('you', INFORM)	109
34.	*drie* ('three')	102
35.	*twee* ('two')	89

Table 8: 35 most frequent words for the 10,000 generated NPs test data

#	Word type	*f*
1.	*dat* ('that')	1937
2.	*te* ('at', 'to', COMP)	1829
3.	*er* ('there')	1353
4.	*niet* ('not')	929
5.	*ik* ('I')	714
6.	*hebben* ('have')	570
7.	*om* ('to')	448
8.	*blijkt* ('turns out')	422
9.	*zijn* ('are', 'be', 'his')	396
10.	*vergeten* ('forgotten')	389
11.	*alle* ('all')	354
12.	*de* ('the', MASC/FEM/PL)	330
13.	*geen* ('no')	330
14.	*ieder* ('every', NEUT)	315
15.	*door* ('by')	304
16.	*elke* ('every', MASC/FEM)	301
17.	*u* ('you', FORM)	294
18.	*elk* ('every', NEUT)	293
19.	*aan* ('to', PREP)	278
20.	*hier* ('here')	276
21.	*jij* ('you', INFORM)	272
22.	*iedere* ('every', MASC/FEM)	268
23.	*daar* ('there')	267
24.	*op* ('on')	259
25.	*veel* ('a lot of')	258
26.	*weinig* ('a few')	250
27.	*je* ('you', INFORM)	249
28.	*massa's* ('masses')	249
29.	*deze* ('this', MASC/FEM)	238
30.	*een* ('a', 'one')	236
31.	*worden* ('become')	233
32.	*werkt* ('works')	221
33.	*wordt* ('becomes')	209
34.	*waar* ('where')	207
35.	*is* ('is')	206

Table 9: 35 most frequent words for the 10,000 generated sentences test data

	NP test experiment	**Sentence test experiment**
Number of word types	3,006	3,442
Number of word tokens	30,129	57,118
Number of hapax legomena	379	411
Number of word types used twice	380	372

Table 10: Lexical statistics of the generated 10,000 NP and sentences test data

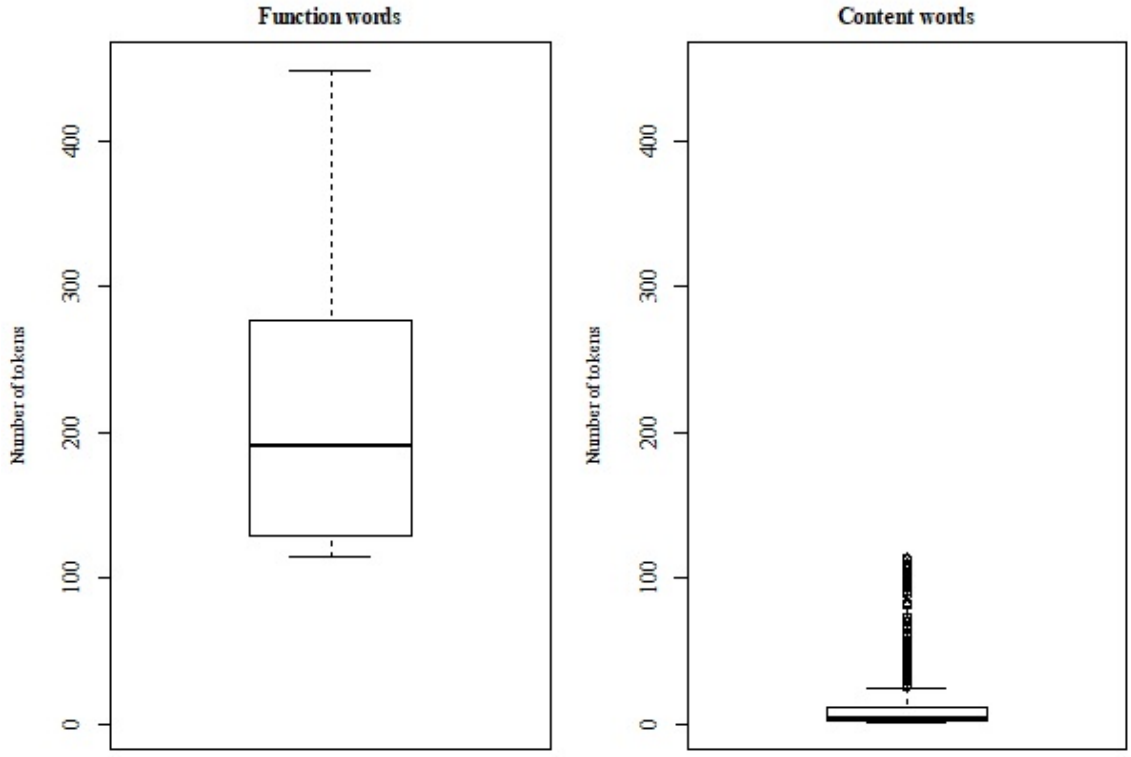

Figure 7: Boxplots depicting function and content words for the 10,000 sentences test data

of 114 tokens. Thus, both test sets provide further support for the suggested parameter settings resulting from the previous experimentations.

Concomitantly, the boxplots corresponding to these data sets visualize the categories resulting from these parameter settings in respectively Figure 6 and Figure 7. Similarly to the results of the training data, it can be observed that for each experiment the groups of function words and content words cluster around different central values. Taking into account these classifications of words, the corresponding lexical entries in the MODOMA daughter grammar are subcategorized as either functional or content categories. Moreover, for the results of the test sets the classifications acquired by the daughter agent have been compared to the grammatical categories employed by the mother language model and Fisher's exact tests have been executed to assess this relationship. Similar to the training experiments, a significant association was found between the categories employed by the mother agent and the categories acquired by the daughter agent, $p < 0.001$ (two-tailed). Thus, the MODOMA daughter agent has learned non-trivial grammatical classifications represented by the currently acquired language model, which it can employ productively to generate novel utterances and analyze data generated by the mother agent.

9. Conclusions and suggestions for future research

This paper discussed how functional and content categories can be acquired by a computational multi-agent language acquisition laboratory involving an adult language model, Delilah, and a daughter agent. The daughter language model has been constructed for the purpose of the MODOMA and employs explicit graph representations of the acquired grammatical knowledge to generate and

parse utterances, which become more specific as a result of language acquisition. Conversely, Delilah consists of a parser and generator that have been independently developed by Cremers and Hijzelendoorn (1995/2024a) to implement a grammar model of Dutch, and employs this model to output and parse utterances (cf. Cremers et al. 2014, for an elaborate discussion of the attributes underlying this language model). Thus, the MODOMA framework requires taking a new approach to research with respect to Delilah: in addition to constructing and improving Delilah to produce utterances and provide parses, the properties of the output of Delilah should be studied as well to enable the daughter to detect the patterns needed for acquisition. To this end, analyses and techniques from for instance the fields of linguistics, psycholinguistics, and corpus linguistics could be applied to gain further insights in the machine-generated output of this language model similarly to how they have been applied to human-generated language data. Taking into account the current rise of applications of large language models, evaluating whether machine-generated language exhibits similar patterns to language produced by humans has become increasingly relevant. This paper provides the results of experiments exploring this approach.

Interestingly, the collection of Delilah utterances generated during the conversation with the daughter agent contains independent exemplars while corpora typically form a cohesive collection of subsequent sentences. Nonetheless, similar patterns corresponding to grammatical phenomena can be detected with respect to both types of linguistic data. This might indicate that these patterns are rather a characteristic of the individual utterances contained in these collections than of the corpora. Crucially, the study presented by this paper reveals that the patterns regarding the divergent frequency distributions of content and functional categories, which are well-established for human language data in the corpus linguistics literature (cf. Powers 1998) are also found in Delilah's output. As Delilah is not configured to reproduce these statistical patterns in its output utterances, the found distribution is a consequence of the specified grammar model.

The initial experiments have been used to set the parameters of the MODOMA acquisition system related to the acquisition of function and content categories, that is, (1) the amount of input data that is required before executing the functional and content categories acquisition procedures, and (2) the threshold for determining whether a word should be classified as either a functional or content category. It was assessed how many input exemplars need to be processed to detect this statistical tendency by conducting experiments on 1,000 and 10,000 NP and sentence training sets. This study revealed that only in data sets containing 10,000 exemplars generated by Delilah a frequency distribution differentiating function and content words can be sufficiently detected. In addition, the threshold value was set to 2 per mil to convert the differences between the distributions of function and content words to discrete linguistic knowledge.

Moreover, these parameter settings have been applied to two test sets containing respectively 10,000 NPs and 10,000 sentences generated by Delilah as well. These experiments confirmed the parameter settings whose values were determined based on the training sets, by similarly allowing for the acquisition of content and functional categories. Figure 6 and Figure 7 visualize the clusters resulting from the acquisition procedures performed on these test sets. Although the input data are noisy and form a continuous distribution, compare Figure 4 and Figure 5, the unsupervised acquisition procedures result in a discrete knowledge-based grammar: each lexical item is classified as either a functional or content category and these properties are specified by feature-value pairs subcategorizing the corresponding lexical graphs in the grammar. Thus, these newly acquired grammatical categories can be used productively to generate and parse utterances by the daughter agent as they specify the combinatorial properties of lexical items during unification.

The combination of properties of the MODOMA system contributes to the field by providing a laboratory approach to modeling language acquisition. Both the adult and child agents interact as integral components of the system. Moreover, the system is fully parameterized, and the acquired grammatical knowledge is explicitly represented and can be consulted. Crucially, as this grammatical knowledge is explicitly represented by the grammar model, the linguistic knowledge and processing of the language model can be straightforwardly consulted by researchers. This design expands the

possibilities for simulating language acquisition as many language models that implement language learning, do not employ explicit knowledge representations. Accordingly, these representations were used to assess whether the grammatical categories employed by the mother language model to generate the input to language acquisition are associated with those acquired by the daughter agent, which is significant for the results of all training and test experiments ($p < 0.001$). This study carried out using the MODOMA supports the research hypothesis that taking into account the differences between humans and artificial intelligence systems the MODOMA framework enables the unsupervised acquisition of discrete grammatical categories such as function and content words by employing statistical techniques. Thus, this result substantiates that a MODOMA can provide a computational laboratory model for simulating language acquisition experiments to study linguistic phenomena.

An important contribution of the study presented in this paper is the validation of the feasibility of a multi-agent computational laboratory for first language acquisition. The application of the principle of Zipf's law thus serves as an essential tool for evaluating the MODOMA approach to modeling language acquisition. Crucially, the experiments designed to acquire function and content categories, using a hybrid approach involving statistical analyses that result in discrete categories, incorporate all the elements found in more complex experiments. As a result, the findings of this study lay a solid foundation for future investigations. For instance, this research paves the way for further studies that can expand and refine the current experiments: future research extending the function and content word experiments could explore the effects of new parameters on the grammar acquired by the daughter agent, the utterances produced by the daughter agent, and the interaction with the mother agent. Examples of such parameters include the number of times the functional and content category acquisition process is executed as well as whether previously acquired categories should be changed when additional data is processed. Another relevant avenue for further study is to investigate the impact of feedback from the mother agent to the daughter agent and its effect on language acquisition, the generated utterances and the interaction between the adult and child agent. Furthermore, the success of the current experiments provides the opportunity to test more complex acquisition procedures. Extending this framework, the acquisition of additional and more intricate grammatical phenomena should be modeled using the MODOMA approach. For example, Shakouri et al. (2025) demonstrate that grammatical categories such as nouns, verbs, and prepositions can be acquired under similar experimental conditions as the function and content words experiment. Building on the results of this study, Shakouri et al. (2025) employ more advanced statistical methods, such as hierarchical agglomerative clustering, to acquire discrete grammatical categories in an unsupervised manner. Thus, the results of this study provide a foundational contribution for the continued exploration and development of this line of research.

Acknowledgements

The authors would like to thank Maarten Hijzelendoorn for his assistance. This paper presents results of the MODOMA project, which was funded by the Dutch Research Council (NWO).

References

Alishahi, Afra and Grzegorz Chrupała (2009), Lexical category acquisition as an incremental process, *Proceedings of the CogSci 2009 Workshop on Psycho Computational Models of Human Language Acquisition*.

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

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

Baldridge, Jason and Geert-Jan M. Kruijff (2003), Multi-modal combinatory categorial grammar, *in* Copestake, Ann and Jan Hajič, editors, *10th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Association for Computational Linguistics, pp. 211–218. https://aclanthology.org/E03-1036.

Balestriero, Randall, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, Tom Goldstein, Adrien Bardes, Gregoire Mialon, Yuandong Tian, Avi Schwarzschild, Andrew Gordon Wilson, Jonas Geiping, Quentin Garrido, Pierre Fernandez, Amir Bar, Hamed Pirsiavash, Yann LeCun, and Micah Goldblum (2023), A cookbook of self-supervised learning, *Computing Research Repository (CoRR)*. https://doi.org/10.48550/arXiv.2304.12210.

Beekhuizen, Barend, Rens Bod, Afsaneh Fazly, Suzanne Stevenson, and Arie Verhagen (2014), A usage-based model of early grammatical development, *in* Demberg, Vera and Timothy O'Donell, editors, *Proceedings of the Fifth Workshop on Cognitive Modeling and Computational Linguistics (CMCL)*, Association for Computational Linguistics, pp. 46–54. https://doi.org/10.3115/v1/W14-2006.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei (2020), Language models are few-shot learners, *in* Larochelle, Hugo, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in neural information processing systems 33 (NeurIPS 2020)*, Vol. 33, Curran Associates, pp. 1877–1901. https://papers.nips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Chaabouni, Rahma, Eugene Kharitonov, Diane Bouchacourt, Emmanuel Dupoux, and Marco Baroni (2020), Compositionality and generalization in emergent languages, *in* Jurafsky, Dan, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, pp. 4427–4442. https://doi.org/10.18653/v1/2020.acl-main.407.

Chaabouni, Rahma, Eugene Kharitonov, Emmanuel Dupoux, and Marco Baroni (2019), Anti-efficient encoding in emergent communication, *in* Wallach, Hanna M., Hugo Larochelle, Alina Beygelzimer, Florence d'Alché Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, Vol. 32, Curran Associates, pp. 6261–6271. https://proceedings.neurips.cc/paper_files/paper/2019/file/31ca0ca71184bbdb3de7b20a51e88e90-Paper.pdf.

Chaabouni, Rahma, Florian Strub, Florent Altché, Corentin Tallec, Eugene Trassov, Elnaz Davoodi, Kory Mathewson, Olivier Tieleman, Angeliki Lazaridou, and Bilal Piot (2022), Emergent communication at scale, *Proceedings of the Tenth International Conference on Learning Representations (ICLR 2022)*. https://openreview.net/pdf?id=AUGBfDIV9rL.

Conner, Michael, Yael Gertner, Cynthia Fisher, and Dan Roth (2008), Baby SRL: Modeling early language acquisition, *in* Clark, Alexander and Kristina Toutanova, editors, *Proceedings of the Twelfth Conference on Computational Natural Language Learning (CoNLL 2008)*, Coling 2008 Organizing Committee, pp. 81–88. https://aclanthology.org/W08-2111/.

Conner, Michael, Yael Gertner, Cynthia Fisher, and Dan Roth (2009), Minimally supervised model of early language acquisition, *in* Stevenson, Suzanne and Xavier Carreras, editors, *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009)*, Association for Computational Linguistics, pp. 84–92. https://aclanthology.org/W09-1112.

Cremers, Crit (2002), ('n) Betekenis berekend, *Nederlandse Taalkunde* **7** (4), pp. 375–395.

Cremers, Crit and Hilke G. B. Reckman (2008), Exploiting logical forms, *in* Verberne, Suzan, Hans van Halteren, and Peter-Arno J. M. Coppen, editors, *Computational Linguistics in the Netherlands 2007 (CLIN 18)*, Vol. 11 of *LOT Occasional Series*, LOT, pp. 5–20. https://www.clinjournal.org/CLIN_proceedings/XVIII/cremers.pdf.

Cremers, Crit and P. M. Hijzelendoorn (1995/2024a), Delilah. Computer software.

Cremers, Crit and P. M. Hijzelendoorn (2024b), Delilah does Dutch. Online demo version. Accessed on 1 October 2024. https://delilah.universiteitleiden.nl/indexen.html.

Cremers, Crit, P. M. Hijzelendoorn, and Hilke G. B. Reckman (2014), *Meaning versus grammar*, Leiden University Press, Leiden. https://doi.org/10.24415/9789087282127.

Dębowski, Łukasz (2021), A refutation of finite-state language models through Zipf's law for factual knowledge, *Entropy* **23** (9), pp. 1148–1182. https://doi.org/10.3390/e23091148.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019), BERT: Pre-training of deep bidirectional transformers for language understanding, *in* Burstein, Jill, Christy Doran, and Thamar Solorio, editors, *North American Chapter of the Association for Computational Linguistics: Human Language Technologies 2019 (NAACL-HLT)*, Vol. 1 (Long and Short Papers), Association for Computational Linguistics (ACL), pp. 4171–4186. https://doi.org/10.18653/v1/N19-1423.

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

Gui, Jie, Tuo Chen, Jing Zhang, Qiong Cao, Zhenan Sun, Hao Luo, and Dacheng Tao (2024), A survey on self-supervised learning: Algorithms, applications, and future trends, *IEEE Transactions on Pattern Analysis and Machine Intelligence* pp. 1–20. https://doi.org/10.1109/TPAMI.2024.3415112.

ter Hoeve, Maartje, Evgeny Kharitonov, Dieuwke Hupkes, and Emmanuel Dupoux (2022), Towards interactive language modeling, *Computing Research Repository (CoRR)*. https://doi.org/10.48550/arXiv.2112.11911.

Lan, Zhenzhong, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut (2020), ALBERT: A lite BERT for self-supervised learning of language representations, *Eighth International Conference on Learning Representations (ICLR 2020)*. https://openreview.net/pdf?id=H1eA7AEtvS.

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

Mandelbrot, Benoit B. (1954), Structure formelle des textes et communication, *Word* **10** (1), pp. 1–27. https://doi.org/10.1080/00437956.1954.11659509.

Matusevych, Yevgen, Afra Alishahi, and Paul A. Vogt (2013), Automatic generation of naturalistic child-adult interaction data, *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 35, pp. 2996–3001. https://escholarship.org/uc/item/2t9414br.

Moortgat, Michael J. (1997), Categorial type logics, *in* van Benthem, Johan F. A. K. and Alice G. B. ter Meulen, editors, *Handbook of Logic and Language*, Elsevier, Amsterdam, chapter 2, pp. 93–177.

Orhan, A. Emin, Vaibhav V. Gupta, and Brenden M. Lake (2020), Self-supervised learning through the eyes of a child, *in* Larochelle, Hugo, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, Vol. 33, Curran Associates, pp. 9960–9971. https://proceedings.neurips.cc/paper_files/paper/2020/file/7183145a2a3e0ce2b68cd3735186b1d5-Paper.pdf.

Pollard, Carl J. and Ivan A. Sag (1986), *Information-based syntax and semantics*, Vol. 1: Fundamentals of *CSLI Lecture Notes*, Center for the Study of Language and Information, Stanford, CA.

Pollard, Carl J. and Ivan A Sag (1994), *Head-Driven Phrase Structure Grammar*, Studies in contemporary linguistics, Center for the Study of Language and Information and University of Chicago Press, Stanford, CA and Chicago, IL.

Powers, David M. W. (1998), Applications and explanations of Zipf's law, *in* Powers, David M. W., editor, *New Methods in Language Processing and Computational Natural Language Learning (N.N., CoNLL 1998)*, Association for Computational Linguistics (ACL), pp. 151–160. https://aclanthology.org/W98-1218.

Radford, Alec, Karthik Narasimhan, Tim Salimans, and llya Sutskever (2018), Improving language understanding by generative pre-training., *OpenAI*. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

Reckman, Hilke 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. 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, Mathieu, Rahma Chaabouni, and Emmanuel Dupoux (2020), "LazImpa": Lazy and Impatient neural agents learn to communicate efficiently, *in* Fernández, Raquel and Tal Linzen, editors, *Proceedings of the 24th Conference on Computational Natural Language Learning (CoNLL 2020)*, Association for Computational Linguistics (ACL), pp. 335–343. https://doi.org/10.18653/v1/2020.conll-1.26.

Schieber, Stuart M. (2003), *An introduction to unification-based approaches to grammar*, Microtome, Brooklin, MA. http://nrs.harvard.edu/urn-3:HUL.InstRepos:11576719.

Shakouri, David P., Crit Cremers, and Niels O. Schiller (2018), MODOMA: Learning in humans and machines (L2HM). Presented at the Learning in Humans and Machines (L2HM) 2018 Conference, École Normale Supérieure, Paris, 5-6 July. http://doi.org/10.17605/OSF.IO/A3RHD.

Shakouri, David P., Crit Cremers, and Niels O. Schiller (2025), Unsupervised acquisition of discrete grammatical categories, *Computing Research Repository (CoRR)*. https://doi.org/10.48550/arXiv.2503.18702.

Steedman, Mark (1996), *Surface structure and interpretation*, Linguistic Inquiry Monographs, MIT Press, Cambridge, MA.

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

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

Steels, Luc (2015), *The Talking Heads experiment: Origins of words and meanings*, Computational Models of Language Evolution, Language Science, Berlin. https://doi.org/10.17169/FUDOCS_document_000000022455.

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

Steels, Luc and Katrien Beuls (2017), How to explain the origins of complexity in language: A case study for agreement systems, *in* Mufwene, Salikoko S., Christophe Coupé, and François Pellegrino, editors, *Complexity in language: Developmental and evolutionary perspectives*, Cambridge Approaches to Language Contact, Cambridge University Press, Cambridge, chapter 2, pp. 30–47. https://doi.org/10.1017/9781107294264.002.

Steels, Luc and Martin Loetzch (2012), The grounded naming game, *in* Steels, Luc, editor, *Experiments in cultural language evolution*, Vol. 3 of *Advances in Interaction Studies (AIS)*, John Benjamins, Amsterdam, pp. 41–60. https://doi.org/10.1075%2Fais.3.04ste.

Steels, Luc, Paul van Eecke, and Katrien Beuls (2018), Usage-based learning of grammatical categories, *in* Atzmueller, Martin and Wouter Duivesteijn, editors, *Preproceedings of the 30th Benelux Conference on Artificial Intelligence (BNAIC 2018)*, Jheronimus Bosch Academy of Data Science (JADS), 's-Hertogenbosch, pp. 253–264.

Takahashi, Shuntaro and Kumiko Tanaka-Ishii (2019), Evaluating computational language models with scaling properties of natural language, *Computational Linguistics* **45** (3), pp. 481–513. https://doi.org/10.1162/coli_a_00355.

Takamichi, Shinnosuke, Hiroki Maeda, Joonyong Park, Daisuke Saito, and Hiroshi Saruwatari (2024), Do learned speech symbols follow Zipf's law?, *in* Ko, Hanseok and Monson H. Hayes, editors, *2024 IEEE International Conference on Acoustics, Speech and Signal Processing 2024 (ICASSP 2024)*, IEEE, pp. 12526–12530. https://doi.org/10.1109/ICASSP48485.2024.10448331.

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

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

Zipf, George K. (1949), *Human Behavior and the Principle of Least Effort*, Addison-Wesley, Cambridge, MA.

Appendix A. Glossary of abbreviations

Abbreviation	Description
COMP	complementizer
FEM	feminine
FORM	formal
INFORM	informal
MASC	masculine
N	noun
NEUT	neuter
PL	plural
PREP	preposition
SG	singular
V	verb

Table A.1: Glossary of abbreviations for grammatical terminology