# Usage of German articles in child language acquisition

**Abstract**:

In general, children omit / use very less articles during their early language development. This is applicable for German language as well. Moreover, the acquisition of German articles is a challenging task during childhood. It is due to the complex nature of German articles. Some of the main reasons for their complexity are gender assignment and high-level syncretism. In gender assignment, one has to learn the correct gender of a noun and select article in accordance with noun’s gender. Also, there are no clear rules about how nouns’ gender can be identified in German. Second, the syncretism of many articles is again making acquisition process more difficult i.e., the same form of articles is used for more than one case, e.g., ‘der’ is the form for both masculine nominative and feminine dative. That’s why children learning German facing challenges in acquiring and using articles. These challenges have led a way to the existence of this project. The main aim of the project is to investigate whether it is really challenging for children to acquire German articles or not by analyzing the overall quantity of articles’ usage during entire childhood. In addition, it also analyzes which articles are used often/less in which childhood stage and shows how articles’ usage are getting transitioned from one stage to another. These analyzes were performed using multiple correspondence analysis (MCA) technique through R programming.

**Introduction**:

In German language, articles play an important role, since they represent the number, gender, and case of a noun[[1]](#footnote-1). The basic forms of article in German can be divided into two major groups i.e., definite and indefinite article. Der (masculine), die (feminine, plural) and das (neuter) are definite. Ein (masculine, neuter) and eine (feminine) are indefinite.

These forms change according to number, gender, and case of a noun. Fig 1.1. shows the entire forms of German article[[2]](#footnote-2).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| indefinite | | | | definite | | | |
| singular | | | | singular | | | plural |
| masculine | | feminine | neuter | masculine | feminine | neuter |
| nominative | ein | eine | ein | der | die | das | die |
| accusative | einen | eine | ein | den | die | das | die |
| dative | einem | einer | einem | dem | der | dem | den |
| genetive | eines | einer | eines | des | der | des | der |

Fig1.1. All forms of German article

The omission of articles in child language development is a well-known phenomenon and there are several literature works about it. As per Clahsen (1984) & Bittner (1997), German-speaking children tend to omit articles in early stages of language development. This project investigates this phenomenon by researching the usage of German articles by children during their language acquisition in childhood. It addresses three main research questions, and they are as follows.

1. At what age, children start to use articles in general? From when they start to use accusative and dative?
2. Which articles' usage is high during childhood: Is it definite or indefinite?
3. What is the overall usage of articles among children, is it more or less (by comparing it with adults' usage)?

In order to analyze these research questions, children ages were divided into five stages in this project as shown in fig 1.2.

|  |  |
| --- | --- |
| stage 1 | 1 to 2 years |
| stage2 | 2 to 3 years |
| stage3 | 3 to 4 years |
| stage4 | 4 to 6 years |
| stage5 | 6 to 8 years |

Fig 1.2. Stages and their corresponding age interval

This study documentation consists of three major sections. The first section talks about the previous literature works related to this project, second section is about the data used in this project, and in third section, those 3 research questions are analyzed with the help of MCA technique using R programming.

**Background literature:**

Some of the main literatures that are related to research questions of this project has been mentioned here.

According to Collings (1990), children started using articles from the age of 2 and half years. At this stage, they mostly use nominative case. They will start using accusative and dative cases from age of 3 and half years.

Mills (1986) stated that children produced indefinite article earlier than definite one and the usage of indefinite feminine article 'eine' is very high until age of 3 years.

According to Kupisch et.al. (2009), when the Mean Length of Utterence (MLU) i.e., no. of words in the utterance is above 3, children (including German speaking kids) use articles in more than 80 percent of all contexts. They have also mentioned that the usage of definite article is more compared to indefinite article.

**CHILDES corpus:**

This project makes use of CHILDES corpus data for its work. Child Language Data Exchange System, otherwise known as CHILDES (MacWhinney & Snow, 1985; MacWhinney, 2000, 2014) is a corpus which contains thousands of children language transcripts across 20+ language (including German). In CHILDES, the majority of the transcripts are from conversations. Childes-db (Sanchez et.al., 2019) is a database formatted mirror of CHILDES in order to improve the accessibility of this corpus. With the help of 'childesr' package in R, childes-db can be accessed and manipulated.

A new dataset has been created from childes-db by filtering only German data. ‘Childes.Ger.article.df’ is the main dataset for this project[[3]](#footnote-3) and fig 1.3. shows sample data of this dataset. It contains data from Caroline, Grimm, Leo, Miller, Rigol, Stuttgart, Szagun, TAKI and Wagner corpora. From this main dataset, a sub-dataset has been created for each research question.

A screenshot of a computer

Description automatically generated

Fig 1.3. Main dataset of this project

**Analyzing research questions using MCA:**

Correspondence analysis is an exploratory space reduction technique for data analysis. It identifies the patterns of association and disassociation in data. It presents result in the form of two-dimensional plot, which shows the relationship between variables in an intuitive manner. Correspondence analysis can be divided into two major classes. They are simple correspondence analysis (CA), and multiple correspondence analysis (MCA). CA is used to analyze the relationship between two variables, whereas MCA is used for analyzing more than two variables. Simple correspondence analysis can be implemented only for contingency table i.e., table with cross tabulation showing the frequency distribution of variables. But, MCA can be implemented for raw dataset and it doesn't need to be converted into contingency table.

For this entire work, MCA has been used, since the dataframe is not a contingency table. In R, there are several packages available for performing MCA. For all the research questions, 'FactoMineR' package (Le et.al., 2008) was used to perform MCA.

**Research question1:**

At what age, children start to use articles in general? From when they start to use accusative and dative?

Implementing MCA:

'childes.Ger.article.df1' is the dataframe for research question1. It contains 1,72,657 rows and 2 columns such as article and stage. Fig 1.4. shows sample data of this dataframe.

A screenshot of a phone

Description automatically generated

Fig 1.4. Dataset of research question1

mca.1 <- MCA(childes.Ger.article.df1)

A screenshot of a computer

Description automatically generated

Fig 1.5. Summary of MCA

Fig 1.5. shows the result of summary. It can be divided into 4 parts such as eigenvalues, individuals, categories and categorical variables.

Eigenvalues represent the variance of data in each dimension. The variance value for first 2 dimensions are always greater than other dimensions. In this result, the first two dimensions together represent 16% of the total variance of data. In general, variance depends on the size of data. When the data size is large, there are chances for the less variance percentage in dimension. Since our data is large, it is not possible to show the entire variance in first 2 dimensions itself. But, it is still recommended to consider, as the first 2 dimensions provide the maximum variance as much as possible.

Individuals part provides details about each row and it is not necessary for this work. Hence, it will not be discussed.

In categories part, cos2 (cosine square) represents the quality of each category. In the first dimension, 'ein' has the highest cos2 value and in the second dimension, 'einer' has highest value. The column 'ctr' provides the percentage of contribution of each category in every dimension. When a category has high percentage, then it contributes the most to their corresponding dimension. In the first dimension, only 'ein' is having high contribution and in the second dimension, only 'einer' is having high contribution.

The last part 'categorical variables' will not be discussed. It contains only the coordinate values of the variables, which is not necessary for this work.

‘Factoextra' is a package which provides an elegant and clear visualization of MCA graph without any overlapping. Fig 1.6. shows the graph of MCA

A graph with red dots

Description automatically generated

Fig 1.6. MCA graph

The column stage and article are variables, and their values are categories. In general, when categories from same variable are grouped together, they are having similar profile, whereas when they lies far apart from each other, they are having. When categories from different variable are grouped together, it means they are closely related to each other, otherwise not in a good relationship.

In fig 1.6., stage1 is separated from stage2. Similarly, stage3, stage4, and stage5 are separated from both stage1 and stage2. But stage3, stage4, and stage5 are somehow associated and there is not much deviation between them.

The articles 'ein', 'einem' are far separated from other categories. 'Die' and 'der' are close to each other, 'das' and 'den' are slightly close to each other.

Relation between stage and article: 'einer' is closely associated with stage1, 'eine' is closely associated with stage2. Likewise, 'den', 'das', and 'einen' are grouped with stage3, also 'dem' is slightly associated with stage3. 'das' lies in between stage3 and stage4 and this implies that 'das' is used in both stage3 and stage4. However, it is very close to stage4. 'der' and 'die' are closely associated with stage5.

Answering research question 1:

At what age, children start to use articles in general?

From the MCA analysis, it is clear that children start using articles at the age of 1 itself. Because, stage1 (1 to 2 years) is associated with an article ('einer'). This is in contrast with the statement of Collings (1990). According to him, children started using articles at the age of 2 and half years.

It means, 'einer' is mostly used by children at this stage (according to this MCA visualization).

From when they start to use accusative and dative?

In German, there is high level of syncretism with article forms i.e., the same form of article is used for more than one case. For example: 'Die' is the form for both feminine nominative and feminine accusative, 'das' is the form for both neuter nominative and neuter accusative. Similarly 'der' is the form for masculine nominative as well as for feminine dative.

The same problem has been mentioned in Collings (1990) as well. "Das for neuter nominative cannot be distinguished from das neuter accusative, just as die feminine nominative cannot be distinguished from die feminine accusative because there is nophonological difference; der, on the other hand, for masculine nominative can easily be distinguished from its morphological counterpart in the accusative, den. Researchers like Clahsen have argued that the emergence of accusative can only be observed when limited to den, the only case marking being visible. As for neuter and feminine, one cannot say for sure if the child uses nominative or accusative" (Collings, 1990).

This consideration was followed in this work as well. So, the usage of accusative will be identified by observing only 'den'/'einen', and dative by 'dem'/'einem', as these forms are the good indicator for accusative and dative respectively.

Coming to the answer, children will start using accusative and dative from the age of 3 years. Because, 'den' and 'einen' are closely clustered with stage3 (3 to 4 years), and 'dem' is also slightly clustered with stage3. This solution is somehow matches with Collings (1990), as he pointed out that the usage of accusative and dative begins at age of 3 and ½ years.

References:

Bittner, D. (1997). Entfaltung grammatischer Relationen im NP-Erwerb: Referenz. *Folia Linguistica*, 31(3–4), 255–283.

Clahsen, H. (1984). Der Erwerb von Kasusmarkierungen in der deutschen Kindersprache. *Linguistische Berichte*, 89, 1–31.

Collings, A. (1990). The acquisition of morphology and syntax in German child language.

Fuchs, J., Domahs, U., & Kauschke, C. (2021). Information structure in language acquisition. Production and comprehension of (in) definite articles by German-speaking children. *Journal of Child Language*, 48(1), 55-87.

Glynn, D. (2014). Correspondence Analysis. Exploring data and identifying patterns. In Glynn, D. & Robinson, J. A. (eds.) *Corpus methods for Semantics. Quantitative studies in polysemy and synonymy,* (pp. 443-486). Amsterdam/Philadelphia: John Benjamins.

Husson, F., Lê, S., & Pagès, J. (2011). *Exploratory multivariate analysis by example using R* (Vol. 15). Boca Raton: CRC press.

Kupisch, T., Anderssen, M., Bohnacker, U., & Snape, N. (2009). Article acquisition in English, German, Norwegian and Swedish. *Little words: Their history, phonology, syntax, semantics, pragmatics and acquisition (Proceedings of The Georgetown University Round Table,* GURT 2007), 223-236.

Le, S., & Husson, F. (2008). FactoMineR: An R package for multivariate analysis. *Journal of Statistical Software,* 25, 1–18.

MacWhinney, B. (2000). *The childes project: The database (Vol. 2).* Psychology Press.

MacWhinney, B. (2014). *The childes project: Tools for analyzing talk, volume ii: The database.* Psychology Press.

MacWhinney, B., & Snow, C. (1985). The child language data exchange system. *Journal of Child Language, 12 (2),* 271–295*.*

Mills, A. E. (1986). *The acquisition of gender. A study of English and German*. Berlin: Springer.

R Core Team. (2017). R: *A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Sanchez, A., Meylan, S. C., Braginsky, M., MacDonald, K. E., Yurovsky, D., & Frank, M. C. (2019). childes-db: A flexible and reproducible interface to the child language data exchange system. *Behavior research methods, 51,* 1928-1941.

1. In addition to articles, adjectives and possessive nouns are also representing noun’s no., gender and case. But, articles are being the basic marker for nouns. [↑](#footnote-ref-1)
2. For this project, only nominative, accusative, and dative cases were considered. But, genitive case and plural forms were not considered in order to make the project clear and simple. Also, children will not use genitive case during their early stages of language development. [↑](#footnote-ref-2)
3. All datasets related to this project can be found at… [↑](#footnote-ref-3)