Science of Law, 2026, No. 3, pp. 20-26
DOI: 10.55284/5kxahw53

Child Speech Genre Classification and their Potential Application in Monitoring Technologies for Safety Purposes

Antonio Rico-Sulayes[1*], Sofía Fernanda Rocha-Velázquez[2]

[1,2]Universidad de las Americas Puebla, Languages Department, Cholula, Mexico.

Keywords:
Conversation,
Corpus linguistics,
Linguistic genres,
Narrative,
Natural language Understanding.

Abstract. In an ever more demanding world, where parents and educators alike struggle to provide attention to young children, the advance of technologies to analyze child speech and monitor its development, along with that of children's themselves, is paramount. This study aims to distinguish two well studied and documented genres in child speech, conversation and narrative. We propose here a straightforward, yet effective method which relies on the comparison of the most common bi-grams, typical two-word sequences. This classifying method is applied to four different corpora, two for Spanish and two for English, with one data set for each of our two targeted genres. Our method uses a semi-automatic extraction technique to obtain bi-grams and successfully compares their distinctive distribution using the statistical test of chi-square. The potential applications of a semi-automated classification of these genres are varied, from child development diagnosis and improvement, to monitoring of children's safety and well-being.

1. INTRODUCTION

It is in childhood that we begin to acquire the necessary skills to produce and understand human language. As this acquisition progresses, we master elaborate language phenomena that allow us to function effectively first in our nearby community and then in society at large. Because child language quickly becomes rich and complex, we can conduct an advanced analysis and exploration of it. Beyond so-called local structures in linguistic production, which include sentences and turns in conversation, language research has also targeted larger structures in organizational terms (superstructures), such as schemata, or sets of patterns and expectations about texts, and larger structures in terms of meaning (macrostructures), such as textual topics and genres (van Dijk, 2011). Representing one of the highest layers of language phenomena, textual genres are types of discourse or texts, such as conversations, news reports, personal emails, among many others. These genres are characterized by a set of typical structures, such as verb tenses or formulaic expressions, and types of activities, defined by characteristics of the context, such as participants and their roles. Here we aim at studying with a novel approach two genres that have been well documented in children's linguistic production: conversation and narratives.

In the context of this research, the differentiation between genres is of particular importance for our study. Therefore, it is worth pointing out that the development of structures characteristic of the genres we have selected follows its own process, alongside that of language acquisition in general. As speakers, we have to manage additional skills besides those that are purely linguistic in order to produce a successful narrative or conversation. The variations concerning the communication intention in different genres offer potentially observable differences in the patterns produced when performing activities such as producing a narrative or participating in a natural conversation.

Nowadays we can analyze language using several computer-based tools that allow us to obtain results about the use of different linguistic forms and their patterns. This type of analysis can be conducted on databases of authentic and reliable, linguistic samples which are available through different corpora. Additionally, with the improvement and advance of technology, this analysis can also be done automatically to provide conclusions that may be useful for the research community and eventually for our day-to-day lives.

In our increasingly technologically-aided societies, many tools are constantly been developed precisely with the purpose of making our lives easier. As the monitoring of children has always been a concern for parents, this context offers an area of opportunity to create or improve technological tools that will assist with keeping track of children's development and the activities they are involved in, particularly during play and other contexts of human interaction.

From a linguistic perspective, while both the acquisition of speech and development of language skills are not a perfect blueprint across all individuals, being able to identify milestones while comparing particular individuals' performance to typical patterns can aid in the early diagnosis of possible impairments. Another potential application of the semi-automatic identification of genres proposed in this study could be the advance of young children's monitoring technologies for safety purposes. The genre detection derived from our analysis could easily be adapted to a monitoring device, so that it can flag the conversation of a child who is thought to be alone, and therefore, the potential interaction with an unexpected interlocutor. This detection could be applied not only to physically present interlocutors, but those remotely reaching the child by either speech- or text-based technological interfaces. In summary, the applications are varied and manifold.

To the best of our knowledge, there is no previous study that has done a comparison of genres in the speech of children as it is presented in this article. Because of this, we have found an area of opportunity, which has rendered the hypothesis in the next section. This hypothesis and its related research questions guide the rest of this study.

2. OBJECTIVE AND HYPOTHESIS

In this research, we have looked into two genres of children's spoken speech: conversation and narratives. We have done it

*Corresponding author.

with the purpose of developing a system that allows us to semi-automatically discriminate these genres in a four-fold corpus with authentic samples of the two mentioned genres for two different languages, English and Spanish. We have intended to find similarities and differences between our target genres through the use of frequent bi-grams, the most typical sequences of two adjacent words, and a statistical comparison of these bigrams across our genres. The hypothesis for this research is that the most common bi-grams for child speech will differ whether the language is being used for conversational speech or narrative. Additionally, this distinction should also be observable and comparable in the obtained bi-grams across the two studied languages. We address the following specific research questions in this study:

- Will there be observable differences between the most common bi-grams obtained from the two distinct genres of conversational speech and narrative?
- How will these differences be explainable through the literature?
- Will these distinctions be equal in nature in Spanish and English?

This study aims at testing our hypothesis and answering the above-listed questions through the following plan. After providing a brief overview of the object of analysis in this research, as we have done so far, in the next section, which presents our theoretical framework, we discuss relevant concepts and background information for the main argument of this study. This framework covers various areas of linguistic inquiry, including language acquisition, the development of narratives, corpus linguistics, genre analysis, and the use of virtual assistants. Then, we present the processes and tools for data collection, and the corpora from which we have derived our databases. In the following section, we present the statistical tools and methods for the interpretation of the collected data. The chi-square test and the use of bi-grams will be outlined and justified in this section. In the next section, the results from the statistical test will be presented for analysis and discussion, offering an interpretation of these outcomes. This section also addresses the practical implications of the results. Finally, in our conclusions we first present a brief summary of the experiments conducted, revisit the hypothesis, and address some of the limitations of our study, as well as propose leads for future research.

3. THEORETICAL FRAMEWORK

In this study we draw on research that has been conducted mainly in five areas of linguistic inquiry: language acquisition, narratology, corpus linguistics, genre analysis, and natural language understanding. The following subsections briefly discuss the concepts from these areas that are especially relevant for our study.

3.1. Language Acquisition

In order to understand the approach to child language used in this investigation, the first concept that needs to be explained is that of language acquisition itself. Human language is different to that of animals, and although the particular characteristics of these differences are still being debated by researchers, we can agree that human language is more complex and sophisticated than the communication methods of animals. What is interesting about human language is that young children develop their linguistic skills naturally and without added effort during their development years. From a linguistic approach and considering nativists' theories, we understand that humans are born biologically programmed for language, which means they have an innate genetic capacity for it. This capacity was defined by American linguist Noam Chomsky as a universal grammar (Centro Virtual Cervantes, 2022). This biological ability develops as the child comes into contact with a particular language and creatively constructs its own version of it, rather than just imitate what other speakers do.

While this hypothesis posits that imitation and repetition are no longer the key concepts leading to language development, interaction with native speakers is still a necessary condition. Participating in conversations allows children to learn about the rules and use of a particular language, as used by those who surround them. As a part of the pragmatic development of speech, conversation develops naturally along with the skills related to it, such as "knowing when and how to take a turn in conversation; how to initiate, elaborate, or terminate a topic; and how to respond to a speaker in keeping with the pragmatic constraints set by the preceding utterance" (Pearson and de Villiers, 2006, p. 688). Among important conversation skills, children have to master the Maxims of Grice which are related to quantity, the amount of information required, quality, the truthfulness of a contribution, relation, the relevance of the information conveyed, and manner, which refers to being clear and easy to understand (Schamberger and Bülow, 2021). To ensure successful communication during a conversation, children must learn to follow these rules. They should also learn to identify when these maxims have been violated and the effects that purposeful violations can bring, such as deceit, humor, and sarcasm or irony (Rowland, 2014).

A key difference between human language and that of animals is that human communication is grammatical because we use linguistic symbols in patterns to convey meaning (Kuhn & Siegler, 2006). This is something that applies to all languages, despite how different they may seem on the surface level. Regarding language acquisition, the earliest evidence of children possessing syntax is shown when children put two words together in an utterance, because, as simple as such an utterance may be, it is already following a set of rules (Berko & Bernstein, 2009). Even in these two-word combinations, categories can already be obtained and distinctions can be made. For example, we may recognize a personal property, which exhibits the relation between a possessor and something that is being possessed, a spatial reference, actions that happen at specific times in relation to people or things, and a discrimination of a specific referent, among other items (López García, 2001). The rules that are followed by a language to put words together in order to form appropriate sentences constitute its syntax.

From the perspective of some linguistics theoreticians, children do not learn syntax, just as they do not learn the language, but rather they obtain the characteristics of it from their environment and map it into a general scheme according to the input that they receive, namely, the elements of their language that the child is exposed to. In this context, a nativist approach means that the child infers the characteristics of words that form the different categories in a language, but also all of the parts that can conform a sentence and how these categories are organized (Liceras & Carter, 2008).

Another important element in nativists' theories is the principles and parameters theory. This theory intends to rationalize the difference in syntactic rules between languages by stating the existence of principles, which specify the basic rules of languages, and parameters, which vary depending on the language (Rowland, 2013). For language acquisition, this means that children have a general structure in their brain, a set of principles, that they fill out according to the characteristics of their native tongue, and the set of parameters. Beyond the combination of turns in conversations and words in syntactical sentences, children need to master the production of longer linguistic structures, such as narratives.

3.2. The Development of Narratives

Longer turns in conversation lead children into their first forms of narration. Stories require a more complex use of language than daily conversations since, besides linguistic knowledge, they represent a task that also relies on genre-specific, structural, and world knowledge (Colozzo, Gillam, Wood, Schnell, & Johnston, 2011). The linguistic knowledge needed to describe an event to someone who has not witnessed such event includes the fact that one must use explicit vocabulary, be clear with pronouns, and effectively use temporal connectors. Furthermore, one must possess cognitive skills that allow us to grasp temporal concepts and cause-effect relationships, as well as reflect and reason about past experiences (Stadler and Ward, 2005). Through a successful narration, speakers manage to plot elements in the narrative while also producing grammatically accurate utterances (Colozzo et al., 2011). Due to all of these skills that need to work together for a successful narration, narratives are an element of language that takes longer to develop but, like the acquisition of language itself, also follows a sequence of levels.

Stadler and Ward (2005) identify five developmental levels for this language function: labelling, listing, connecting, sequencing, and narrating. Each level requires and exhibits more complex language skills than the previous one. Children begin with the use of nominal labels and repetitive syntax, then manage topic-centered lists of attributes or actions of a character, move on to a central topic with character actions that are linked to related characters or events, progress when they are able to use consistently correct temporal sequencing and cause and effect, and finally include all of the components of the previous levels, as well as develop plots with evidence of planning to reach goals.

The previous explanations show a conception of universality across languages. In this view, language acquisition is possible from the very early stages of development regardless of the language learnt, which leads us to believe that all human languages might share underlying features that allow for this to happen. This view supports Chomsky's universal grammar theory, which essentially states all languages although different on the surface share a set of grammatical principles and a set of parameters that specify which aspects of grammar can be set to different values depending on the specific language (Rowland, 2014).

3.3. Corpus Linguistics

In this study, we will conduct a number of experiments using corpus linguistics tools and procedures. Corpus linguistics is "an empirical methodology that employs a large, systematically organized body of natural texts (the corpus) to analyze actual patterns of language use" (Rutherford, 2005, p.354). Furthermore, a corpus is a large, electronically stored collection compiling naturally occurring examples of language (Bennerr, 2010). In this way, researchers working with corpora can access numerous real-life examples of language through which analysis can be performed in order to test a hypothesis and include empirically grounded data. A corpus-based analysis of language generally includes the following four features (Reppen & Simpson-Vlach, 2014): (1) It is empirical, which means that it analyses the patterns of natural texts. (2) It uses what is called a large and principled collection of natural texts for this analysis. (3) Computers are used for analysis along with automatic techniques. (4) Analytical techniques are both quantitative and qualitative.

As just described, a corpus-based methodology effectively makes use of technological advancements in order to explore and analyze naturally occurring language. However, this is not to say this approach has not faced criticism. Chomsky, for instance, has famously objected to corpus-based work. From these objections, it is valuable to mention two: the statement that, despite their size, corpora can never be comprehensive and fully account for all possible uses and the argument that linguistic importance is not necessarily reflected in the frequency of use (Rutherford, 2005). It is possible to see the problems that these objections arise from since language is undoubtedly too large to be completely sampled and false conclusions can arise from a wrongful generalization of results. Nonetheless, this does not apply solely to corpus linguistics, as the same could be argued about other methodologies and research methods. Even so, since these two objections were directed at the practicality, and consequently the validity of results obtained from this methodology, corpus linguistics aimed to develop solutions, such as the development of larger corpora in order to encompass general linguistic features in usage. This was done, of course, while also maintaining smaller more specific corpora, as well as increasing use of statistical tools and approaches with external referents, in order to minimize sampling bias and context-bound uses.

The advancements in computational software and tools have also facilitated the use of corpora in research by making the acquisition of results faster, more efficient and precise. Although the statistical techniques can be made fully automatic, human researchers are the ones deciding what topics are worth looking into, what information to extract from a corpus and how to interpret the results offered by the computer. For this, both quantitative and qualitative techniques are necessary for the proper analysis of corpus linguistic research (Reppen & Simpson-Vlach, 2014).

3.4. Genre Analysis

Genre is a term that has had various characterizations in the context of literature studies, but also within linguistics. This has created a multi-faceted concept that produced various approaches for analysis, as well as an elusive definition. Mauranen (1998) points out two main perspectives in the study of this phenomenon: the sociocognitive approach and the one used within corpus studies. The former understands genre as a social, dynamic, and interactive process, while the latter sees it as a large label to encompass variation in discourse types. Understandably, the variation in the definition and approach to genre is motivated by the purposes of the analysis (Bhatia, 2002). In general terms, Bhatia defines genre analysis as "the study of situated linguistic behaviour" (2002, p. 4). Besides this definition, there are many other with slight variations. What we can extract from this variability is a change in the approach to defining genres, because while traditionally authors did try to give in their definition a limited number of properties, nowadays the focus is more broad, open, and fluid (Rutherford, 2005). Additionally, this broadness is also followed by levels of abstraction and the argument of a hierarchical approach, which then means that instead of having a single level at a specific time and place, one can recognize specific genres within a broader class if necessary.

For corpus studies, text-internal characteristics, rather than external criteria such as topic, medium, and authorship, have become increasingly important for the analysis of corpora (Mauranen, 1998). What this means is that automatic analytical and statistical procedures are applied in order to extract patterns that consider the presence of various features to indicate the genre. This approach contrasts with and is more informative than an analysis of individual features one at a time (Mauranen, 1998). The danger of such an approach could be to forget precisely that social focus, by assuming a systematic standardization of language (Mauranen, 1998). However, the focus on many specific individual textual features for the definition of different genres especially suits automated approaches for their detection, like the ones used in computational linguistics, and semi-automated ones, like the

one we will rely on in this study.

3.5. Virtual Assistants and Natural Language Understanding

The final concept needed as background for this research is that of virtual assistants and the increasing interaction between computers and humans. The use of intelligent virtual assistants (IVAs) has increased exponentially in recent years due to the fast development of technology. As the use of IVAs continues to grow, so do the number of tasks we expect them to perform and the number of interactions we have with them, which we also expect to be of a greater quality. The way in which IVAs process language is the relevant aspect to explore in this section. Different characteristics of language have been mentioned above, and while its understanding does not represent a great challenge to humans, the same cannot be said for computers. The complexity of human language represents a challenge for computers and while research in the areas of speech recognition and language processing has existed for decades, it is only in more recent times that the applications for these topics have spread to more quotidian systems rather than remaining in laboratories (Bates, 1995). It is therefore common now for all IVAs to have a component for natural language understanding (NLU), which "maps user inputs, or conversational turns, to a derived semantic representation commonly known as the intent, an interpretation of a statement or question that allows one to formulate the best response" (Beaver and Mueen, 2021, p. 30). Within NLU there is the collection of grammar rules, syntax and semantics that allows the system to map input language and defines how it does it. This collection is a language model and can both be trained through machine learning or be manually constructed (Beaver and Mueen, 2021). Naturally, since language is a fluent, ever-evolving human resource, language models cannot be simply created and ready for any situation. To improve them and ensure the quality of IVAs interactions and responses, they need to be continuously reviewed, a task that may include adding new vocabulary or rules, or revising existing ones to correct inaccurate mappings in the model (Beaver and Mueen, 2021).

Finally, we have mentioned IVAs because we believe this could be one of the most relevant applications of our genre detection model. As we mentioned before, IVAs tasks have been increasing constantly in recent years, including tasks in domains as diverse as the travel domain, where they may help a costumer to book a trip, the routing domain, in which they act as call center agents guiding users through menus, and the tutoring domain, where they help students and trainees to learn or review course contents (Jurafsky and Martin, 2026). The very last domain, the educational one, is where we think a genre detection model for young children could be implemented for various purposes, such as diagnosis of early language acquisition issues.

This chapter reviewed the framework necessary to provide the tenets and principles behind language acquisition in general, the role of both conversation and narratives in this human enterprise, the use of large corpora of natural language for the study of these types of phenomena, and how the use of these corpora matches with the study of linguistic genres. At the end, we have also provided a background for the connection between this study and the ever-growing applications of IVAs, particularly in its tutoring domain. Our review of the theory of language acquisition and genre analysis should also provide a justification for the differences and similarities between the two languages that have provided the data for this research, English and Spanish. Differences will be particularly important in the language structures used for narration and conversation, thus permitting their discrimination and detection, but similarities across languages allow us to succeed in our task in both languages with a common approach. Therefore, the data will show that the word patterns that repeat themselves most in each genre will be distinct, yet observable in another language.

4. DATA

The data analyzed was extracted from four corpora, two for Spanish and two for English. These corpora are available for use on the TalkBank project site under the Child Language bank of CHILDES. Each corpus was created as part of a different study and because of this, the number of participants, their characteristics, and the topics in the data collection are different. The features of each corpus will be described below according to the information provided on the source site along with the transcriptions.

The English corpora both consist of transcriptions of the speech of North American English-speaking children. The corpus for narrative addresses three topics for the participants to talk about: McDonald's, being late for school, and aliens. The corpus includes data from 250 target and 520 control participants in an age range of 5 to 11 years old. For this research, we used the data from 20 participants for each age group, 10 males and 10 females, which in total comprised the transcripts from 140 participants. Since language impairment information was provided in the labeling of the corpus data, it is relevant to note that only narrations of kids without speech impairment were selected to be used in this research.

The second English corpus, which had a conversational approach, was created in 1973. It contains 46 files of children in an age range going from two to five years old. The conversational data in this corpus was recorded while two children played in a room by themselves, that is without the intervention of the researcher or other adults. We used all 46 transcriptions in this case.

In the case of Spanish, the corpus for narrative includes the data from 24 children, 8 from each of the three age groups of six, nine, and twelve. The children were prompted with 5 different tasks during two separate sessions. Although all of the tasks they were given were narrative-oriented, not all of them were useful for the purpose of this research. Only transcripts from the second and fourth tasks were selected, because it was during these tasks that the participating children were asked to produce their own narratives, instead of judging and commenting on the ones given by the instructors, as was the case for the other three tasks, which we decided not to include.

Finally, in the Spanish corpus containing children's conversational speech, the participants were between four to twelve years of age. Originally the corpus contains 81 transcriptions, however, for the data to better match that of its English counterpart, we only used the four transcript files in which the children were interacting with each other in small groups, instead of including the rest where they were doing it individually with the researcher.

The main limitation of the data used in this study stems from the fact that they are from different research works which followed different methods to collect the data. We decided to use these various resources, because this was the most practical way to produce a rich enough database with the necessary genre distinctions and amount of samples.

5. METHODOLOGY

As has been mentioned before, this research is based on the data obtained from four different corpora, two for Spanish, and two for English. There is one corpus for each of the two genres, conversational and narrative, in both languages. We processed and analyzed these various databases through a corpus linguistics approach, which allowed us not only to have large amounts

of naturally occurring data, but also handle it for analysis purposes. The corpora provide complete transcriptions of the audio files recorded for all the participating children. These transcriptions were pre-processed, cleaned of annotations and interventions from the researchers, in order to maintain only the children's speech before being analyzed.

The data extraction was then done semi-automatically with the use of AntCont version 5.3.8, a corpus analysis software that examines text and finds concordances (Anthony, 2019). This software was used to obtain the most common bi-grams in each of the corpora as well as their frequency. Bi-grams are a form of n-grams which are sequences of words, namely bi-grams are sequences of two adjacent words. N-grams have been used to create language models, calculating probabilities of words based on previous words and producing probabilities for longer sequences (Jurafsky & Martin, 2026). In other words, the semi-automatic extraction provided us with the most common two-word sequences in each corpus. These resulting bi-grams were compared across both topics with the use of chi-square to determine the statistical significance of their distribution. Chi-square is a statistical method that allows for this evaluation by comparing the observed and expected values of the data in a context. Therefore, this tool provided statistical evidence of the observed differences in the results in order to determine whether the most frequent bigrams in one genre were associated with it in a distinctive way, when compared to their frequency in the other genre, so that this association reached statistical significance.

Subsequently, to determine the similarity of the resulting bi-grams across languages, we compared the results obtained from English and Spanish with the ones in the corpus of the same genre in the other language. It would not be possible to use the same statistical method to compare different languages, since the probability of finding English or Spanish using a corpus in the opposite language would be nil. Consequently, this part of the analysis to compare if the most frequent bi-grams show similarities across languages is qualitative instead of quantitative. To judge these similarities, we have compared the bi-grams based on the function these word sequences serve in the context of the utterances.

6. RESULTS

In order to perform the mathematical calculations explained before, first, the data obtained from each corpus will be presented. As it was mentioned, the first analysis step in the methodology was to input the cleaned transcriptions into AntConc to obtain the ten most common bi-grams in each genre for each of the two languages targeted. Tables 1 and 2 show the results for each language in descending order, starting from the most common n-gram, along with the frequency with which they were obtained.

Table 1. Most common bi-grams in the English corpora.

Narrative		**Conversation**	
Bi-gram	**Frequency**	**Bi-gram**	**Frequency**
and then	253	this is	212
and they	176	going to	190
he was	160	do you	185
and he	154	I don't	184
and the	132	in the	171
and a	120	I'm going	127
they were	112	you can	117
to go	98	in here	105
to the	91	let me	99
it was	88	I know	95

For the results from the English corpora, it is important to mention that contractions were considered as one word because they are treated as such in spoken language. This can be observed in the fourth and sixth most common bi-grams in the conversation corpus.

Table 2. Most common bi-grams in the Spanish corpora.

Narrative		**Conversation**	
Bi-gram	**Frequency**	**Bi-gram**	**Frequency**
y ya	202	es que	66
a la	167	que no	58
y se	146	en la	43
en la	128	en el	41
y que	125	a la	40
ya no	117	ha sido	40
es que	107	de la	37
y me	104	que es	37
a mi	99	que se	37
y le	88	a ver	36

For the Spanish narrative results, we did not include *y y,* 'and and' because it didn't serve a function in speech and was rather a stuttering hesitation when speaking. We also omitted *mi papa,* 'my dad' and *mi mama,* 'my mom' because these were specific to the topic of the narrations and would not represent the language in general. For this same reason, we also omitted *la guerra,* 'the war' from the results of the conversation corpus. With the elimination of these four cases, all bi-grams used exhibited mostly syntactical functions, rather than content related information.

After obtaining the most frequent bigrams, we cross-compared the 5 most common bigrams of each corpus with the 5 corresponding ones in the opposite corpus with the use of chi-square. For this, we had to extract the frequency with which the bi-grams occurred in the other genre for the same language. The chi-square calculations were done with Excel. The total chi-square values for each pair are presented in Table 3, below.

Table 3. Chi-square total value for each bi-gram pair compared in each language.

English		Spanish	
Bi-grams compared	**Chi-square value**	**Bi-grams compared**	**Chi-square value**
and then, this is	345.56	y ya, es que	62.60
and they, going to	306.89	a la, que no	24.77
he was, do you	270.28	y se, en la	8.70
and he, I don't	267.38	en la, en el	7.74
and the, in the	113.81	y que, a la	18.38
and a, I'm going	179.76	ya no, ha sido	128.70
they were, you can	217.33	es que, de la	1.79
to go, in here	124.96	y me, que es	66.65
to the, let me	85.38	a mi, que se	3.63
it was, I know	136.94	y le, a ver	29.72

7. ANALYSIS AND DISCUSSION

For chi-square, in a comparison like the one performed here, where there are two independent variables (narrative and conversation) and two dependent variables (two of the most frequent bi-grams), the degree of freedom (DF) equals 1. This is because DF equals the number of independent variables minus 1, times the number of dependent variables minus 1 (Lindquist & Magnus, 2018). Also, since the significance level in linguistics needs to be at least 0.05, or $p<0.05$, we need a chi-square score of at least 3.84 for the results to be significant. For English, all of the results rendered a considerably higher value, therefore all of them are statistically significant to a very high level of $p<0.001$. This means that the obtained bi-grams for this language have been considerably different depending on the genre and that therefore their appearance does depend on or is strongly associated with it.

Spanish presented a slightly different situation, as it can be seen of the right half of Table 3 above. In general, the extracted results have been similar, with some bi-grams appearing on the most frequent list for both genres, such as *a la,* 'to the' and *en la,* 'in the'. However, for the Spanish comparison of bi-grams, two of the chi-square results have not been statistically significant, and the score values of this statistical test in general have been lower compared to the English results. The only noticeable exception is the case of *ya no, ha sido,* 'not anymore, has been' which is the pair with the highest chi-square value in Spanish. This is not too surprising, considering *ha sido* did not have any concordance hits in the narrative corpus and *ya no* only had 8 in the conversation corpus. It is also important to mention the last pair, *y le, a ver,* 'and to him/her, let's see', since *a ver* did not have the same meaning for both corpora. In the one regarding conversation, it was used most commonly as an expression to draw the attention of others before speaking, while in the narrative corpus it is used as the verb meaning 'to see'. Therefore, their use is not equal, and the real chi-square value would be different if one were to extract only those with the same meaning. Taking all of this into account, we could conclude that Spanish structures do not vary as much as English structures do depending on the genre the speech has been produced for. However, a distinctive distribution of bi-grams can be observed in both English and Spanish.

Comparing the two languages, on the narrative corpora results, the bi-grams do seem similar. In English, 5 of the results begin with *and* while in Spanish the same number do with its equivalent *y*. Both also include the use of pronouns, English uses the singular and plural for the third person, while Spanish has third and first-person singulars. We have presented these results as comparable because they serve the same narrative purpose in both languages, the difference in person is caused by the topic of the narrations because while the ones for English all involved fictional made-up stories, the ones for Spanish included both fictional and personal stories.

The results from the conversation corpora also contain comparable results among the most common bi-grams. For instance, *en la* or *en el* from the corpus in Spanish have the same meaning as *in the* from the English one and showed up at a similar position in the frequency list. The 10th result *I know* and *a ver* is similar, in the sense that in both cases they are used as connecting expressions, rather than having the literal meaning of their verbs.

These results, which have proven to be significant, could be applied in machine learning contexts of natural language understanding, as in the production of a tutoring-oriented IVA. Providing software with data from each genre can be used to train an agent or application to identify what a child is doing in terms of linguistic production over a specified period of time. An accurate prediction could give parents or educators a general evaluation of children's performance and level of development.

8. CONCLUSIONS

The results of this research provided statistical and qualitative evidence to prove the hypothesis. The most common bi-grams for child speech differ whether the language is being used for unprompted conversation or narrative. This has been proved to the point that different frequency distributions for these linguistics items have shown to be different and specific to the genre reaching statistical significance. We can see that the most common bi-grams in the narrative corpora are narrative structures used to list sequences of events in a story. On the other hand, the most common bi-grams in the conversation corpora are first-person structures and those used to refer to someone or something else. When comparing the languages, the narrative and conversation genres give evidence of similar results. However, we believe that one of the reasons why the English results were more statistically significant might be due to the fact that the same varieties of Spanish were not used in both corpora for this language. The conversation corpus had participants from Spain while the narrative corpus participants were Mexican. Because of this, the language used was not as closely resembling as it could have been, as was the case for the English corpora, which both were created in the United States.

Further research could provide more evidence for the former explanation. Data that does not have these cultural differences could give different statistical results. Furthermore, perhaps a comparison across more languages could also prove valuable for the qualitative aspect of the research in demonstrating comparable results for different languages. Additionally, with enough available corpora to provide data, this research could be expanded to compare the statistical results from typical speech development and that of children with speech impairment. A variation in results could then be used to form a database to train software for the early detection of linguistic impairment. Following our reasoning at the beginning of this article, we also believe that an effective identification of genres, as the one achieved here, could be used in young children's monitoring technologies for safety purposes. In summary, the extensions and implications of our work are manifold and we hope to carry some of these in future research.

REFERENCES

Anthony, L. (2019). *AntConc (Version 3.5.8)* [Computer software]. Waseda University. https://www.laurenceanthony.net/software

Bates, M. (1995). Models of natural language understanding. *Proceedings of the National Academy of Sciences, 92*(22), 9977–9982. https://doi.org/10.1073/pnas.92.22.9977

Beaver, I., & Mueen, A. (2021). On the care and feeding of virtual assistants: Automating conversation review with AI. *AI Magazine, 42*(4), 29–42. https://doi.org/10.1609/aaai.12024

Benedet, M., & Snow, C. (2004). *CHILDES database Spanish BecaCESNo corpus*. https://doi.org/10.21415/T5ZG77

Bennett, G. (2010). *Using corpora in the language learning classroom: Corpus linguistics for teachers*. University of Michigan Press.

Bhatia, V. K. (2002). Applied genre analysis: A multi-perspective model. *Ibérica: Revista de la Asociación Europea de Lenguas para Fines Específicos*, (4), 3–19.

Berko Gleason, J., & Bernstein Ratner, N. (Eds.). (2009). *The development of language* (7th ed.). Pearson.

Centro Virtual Cervantes. (2022). Hipótesis innatista. Centro Virtual Cervantes. https://cvc.cervantes.es/ensenanza/biblioteca_ele/diccio_ele/diccionario/hipotesisinnatista.htm

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Colozzo, P., Gillam, R. B., Wood, M., Schnell, R. D., & Johnston, J. R. (2011). Content and form in the narratives of children with specific language impairment. *Journal of Speech, Language, and Hearing Research, 54*(6), 1609–1627. https://doi.org/10.1044/1092-4388(2011/10-0247)

Garvey, C., & Hogan, R. (1973). Social speech and social interaction: Egocentrism revisited. *Child Development, 44*, 562–568.

Gillam, R. B., & Pearson, N. A. (2004). *Test of Narrative Language*. Austin, TX: Pro-Ed.

Hess Zimmermann, K. (2003). *El desarrollo lingüístico en los años escolares: Análisis de narraciones infantiles* (Unpublished doctoral dissertation). El Colegio de México.

Jurafsky, D., & Martin, J. H. (2026). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (3rd ed.). Draft manuscript. https://web.stanford.edu/~jurafsky/slp3/

Kuhn, D., & Siegler, R. S. (Eds.). (2006). *Handbook of child psychology* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

Liceras, J. M., & Carter, D. (2008). La adquisición del léxico. In E. de Miguel (Ed.), *Panorama de lexicología* (pp. 371–404). Barcelona: Editorial Ariel.

Lindquist, H., & Magnusson, L. (2018). *Corpus linguistics and the description of English* (2nd ed.). Edinburgh University Press.

López García, Á. (2001). Sintaxis mínima. *Revista de Investigación Lingüística, 4*(1), 97–108.

Mauranen, A. (1998). Another look at genre: Corpus linguistics vs. genre analysis. *Studia Anglica Posnaniensia*, 303–315.

Pearson, B., & de Villiers, P. (2006). Discourse, narrative and pragmatic development. In K. Brown (Ed.), *Encyclopedia of Language & Linguistics* (2nd ed., pp. 686–693). Elsevier. https://doi.org/10.1016/B0-08-044854-2/00841-5

Reppen, R., & Simpson-Vlach, R. (2019). Corpus linguistics. In N. Schmitt & M. Rodgers (Eds.), *An introduction to applied linguistics* (pp. 91–108). Routledge.

Rowland, C. (2014). *Understanding child language acquisition*. Routledge.

Rutherford, B. A. (2005). Genre analysis of corporate annual report narratives: A corpus linguistics-based approach. *Journal of Business Communication, 42*(4), 349–378. https://doi.org/10.1177/0021943605279244

Schamberger, C., & Bülow, L. (2021). Grice and Kant on maxims and categories. *Philosophia, 50*(2), 703–717. https://doi.org/10.1007/s11406-021-00398-4

Stadler, M., & Ward, G. (2005). Supporting the narrative development of young children. *Early Childhood Education Journal, 33*(2), 73–80. https://doi.org/10.1007/s10643-005-0024-4

van Dijk, T. A. (Ed.). (2011). *Discourse studies: A multidisciplinary introduction* (2nd ed.). Thousand Oaks, CA: SAGE Publications.