**Early language acquisition across language families**

Cognitive Science, University of Aarhus  
Jens Chr. Skous Vej 2, 8000 Aarhus, Denmark

# Abstract

Bleses, et al. (2008) argue that children across languages do not acquire language in the same manner. Different word segmentation cues exist between languages, and several factors such as word stress patterns or grammatical rules can be contributed to the language family a language is a part of. This means that there might be differences in language acquisition between language families. On this background, this paper will argue not only that a difference in language acquisition is present between language families, but that children will acquire a Slavic language faster than a Romance and Germanic language. Furthermore, the paper will argue that children will acquire a Romance language faster than a Germanic language. Through analysis of CDI data from WordBank, evidence is presented displaying that there in fact is a difference in language acquisition between language families, however not exactly as the hypotheses propose: Slavic learning children are faster at comprehending language than Germanic and Romance learning children, but Germanic children overall acquire language faster than Romance children.

Keywords**:** CDIs, Indo-European language family, language acquisition, language families

# Introduction

In previous research, it has been found that infants do not acquire language in the exact same way across languages (Bleses, et al., 2008; Lleó & Demuth, 2015). Some factors influence the infants’ language learning trajectory. But what must some of these factors be? There are several proposals on this area – the parents’ education, the different phonology of languages or the child’s gender. This paper will argue that the language family of which the language is a part of will be a factor.

Werker (1995) suggests that infants have developed sensitivities through evolution making them “universal listeners” capable of learning any language in the world. However, infants stop being “uni-versal listeners” within the end of their first year of life, as they develop an enhanced sensitivity to the linguistic properties of the language they are acquiring. In order for infants to acquire language, they must first learn word segmentation: parsing fluent speech into single words. This becomes possible when infants are able to detect and exploit cues to word boundaries, and this ability develops in the last half of infants’ first year. It has been found that English learning infants develop this word segmentation ability when they are between 6 and 7.5 months of age (Johnson & Jusczyk, 2001).

Several cues can be used in word segmentation, and it is important for infants to develop a sensitivity to these cues in order to be able to segment words. Some of these cues include prosodic markers such as word stress patterns, and phonotactic constraints, which are restrictions on the possible order of sounds within a word. For instance, the consonant cluster /mg/ cannot begin a word in English, but /gl/ can. Studies indicate that infants between the age of 6 and 9 months develop sensitivity to two distinct word segmentation cues: phonotactic patterns and word stress patterns. Segmentation by word stress patterns would seem like a good strategy for infants, as approximately 90 % of English words with lexical meaning follow a trochaic structure with a strong-weak stress pattern, such as the words “habit” and “working” (Cutler & Carter, 1987), and as Slavic languages such as Czech, Slovak and Hungarian have word-initial fixed stress, meaning stress always falls on the first syllable of a word (Hanulikova, 2009).

English-learning infants at the age of 7.5 months have been observed to correctly segment bisyllabic words with trochaic strong-weak structure, but they are not able to correctly segment weak-strong structured words until the age of 10.5 months. Jusczyk (1999) suggests that another segmentation cue that infants develop sensitivity to around the age of 10.5 months is the allophonic cues within words. An allophone is a phonetic variant of the same phoneme. For instance, the aspirated /t/ in “talk” and the unaspirated /t/ in “stalk” are allophones of /t/. The correct segmentation of strong-weak words at the age of 7.5 months suggests that English-learning infants develop sensitivity to prosodic cues such as word stress patterns first, and then develop sensitivity to other word segmentation cues as they age, such as the allophonic cues. This is consistent with what have been found in previous research, which shows that infants from birth are sensitive to prosodic cues such as word stress patterns (DeCasper & Spence, 1986). (Johnson & Jusczyk, 2001)

According to Harley (2010), infants start babbling around the age of 6 months and display signs of comprehension around the age of 8 to 9 months, responding appropriately to what they hear. At around 18 months of age, children go through a phase called “the vocabulary explosion”. Just before this phase, children produce holophrastic speech, which is when a single word or sound stands for something more complex. During the vocabulary explosion, a drastic improvement in the children’s comprehension happens and they start to combine words in their production in a telegraphic manner. The telegraphic speech is when speech is simplified but still conveys meaning, such as “mommy hug” implying that the child wants a hug from their mother. The reason for this vocabulary explosion is that children around the age of 10 month start “bootstrapping”, which is when the child knows some words and uses the knowledge of these words to figure out what other words mean. As time progresses, the child knows more and more words and can figure out the meaning of more and more words.

According to Dale & Goodman (2005), grammatical development and vocabulary production are strongly related. Following the idea of prosodic cues as a word segmentation strategy, Gerken (1996) has found that English learning children’s early speech production will include grammatical morphemes such as determiners with trochaic strong-weak structure, but not grammatical morphemes with weak-strong structure. Moreover, Spanish learning children produce trisyllabic words and nouns with determiners much sooner – almost half a year – than German learning children. At the age of 16 to 17 months, Romance language learning children begin to produce proto-determiners, whereas Germanic language learning children do not start to produce proto-determiners until the age of 22 months. This difference might be determined by how the determiners are prosodified in the given languages. (Lleó & Demuth, 2015)

In a CDI based study, Bleses et al. (2008) compared 17 studies and found that Danish learning children score low on vocabulary comprehension and production in early stages of language development. They also found that the median Mexican Spanish learning child comprehends ten times as many words as the median Swedish learning child at the age of eight months, and that the median Croatian learning child comprehends more than twice as many words as the median Danish learning child at the age of 15 months. Furthermore, at the age of 15 months, the median Spanish learning child produces twice as many words as the median Danish learning child, and the median Croatian learning child produces three times as many words as the median Danish learning child. The great difference between these languages is quite appreciable.

The languages mentioned above can be split into three different language families within the Indo-European language family: Germanic, Romance and Slavic. The Germanic language family includes languages such as Danish, English, German, Swedish, Dutch and Norwegian. The languages are derived from Proto-Germanic or Common Germanic and can be subdivided into North, West and East Germanic. English and German are West Germanic languages, whereas Danish and Norwegian are North Germanic languages. Today, all the East Germanic languages, such as Gothic and Vandalic are extinct. The Germanic languages have several unique linguistic characteristics in common, such as the sound changes known as Grimm’s Law and Verner’s Law affecting stop consonants and introducing fricatives (e.g. what used to be /t/ in Proto-Indo-European became the aspirated /θ/ in Germanic), word-initial stress on most words resulting in many monosyllabic words, and a large number of vowels. Furthermore, in Germanic languages, the grammatical rule of verb-second order is followed (e.g. the Danish “Om morgenen **drikker** jeg te” and the German “Morgens **trinke** ich Tee”), although this is only the case in certain wordings in English. (Germanic Languages, n.d.)

The Romance language family includes languages such as Spanish, French, Italian, Portuguese and Romanian. The languages are derived from Italic, which in turn is derived from Vulgar Latin. The languages are difficult to subdivide, as the languages exist rather as a continuum than as distinct categories. Some of the shared characteristics of the Romance languages include predictable word stress patterns, fairly strict subject-verb-object word order and null subject clauses resulting from verb inflection (e.g. the Spanish “Está cansada: (She) is tired”. However, the case of null subject clauses is not present in French. (Romance Languages, n.d.; Rudder, n.d.)

The Slavic language family includes languages such as Russian, Slovak, Croatian, Hungarian, Polish, Slovenian and Czech. The languages are derived from Balto-Slavic and can be subdivided into East, South and West Slavic. Russian is an East Slavic language, whereas Slovak and Czech are West Slavic languages, and Slovenian is South Slavic. The Slavic languages vary in prosody, but have several common characteristics such as palatalizations (e.g. phonemes such as /ch/ and /zh/), no use of articles (except in Bulgarian and Macedonian), and word-initial fixed stress in most language. (Wagner, n.d.; Slavic Languages, n.d.)

Based on this knowledge and especially the studies conducted by Lleó & Demuth (2015) and Bleses, et al. (2008), this paper will argue that a) children will acquire a Slavic language faster than children acquiring a Germanic or Romance language, and that b) children will acquire a Romance language faster than children acquiring a Germanic language. This implies that the infants’ age when acquiring their first word(s) will happen in an orderly manner from Slavic language learners first, then Romance, followed by Germanic. Furthermore, the rate at which language is acquired and the end acquisition score at the age of 30 months will be ordered in the same manner. As the comprehension and production of language do not happen at the same time (Bleses, et al., 2008), language acquisition will be split into these two categories.

# Materials and methods

## Data

The data are collected from WordBank, which is an open database of children’s vocabulary development. This database contains data from the MacArthur-Bates Communicative Development Inventories (CDIs), which are information about children’s early language development reported by their parents.

The data used in the analysis include information from Danish, English, Norwegian, French (Quebec), Spanish, Italian, Russian, Croatian and Slovak learning children of the age of 8-30 months, meaning that each language family is represented by three languages. It includes information about the infants’ age, gender, language, maternal education (for some of the languages) and a comprehension and production score. A score of 1 is equal to the comprehension/production of one word. Swedish, German, English (British) and Czech data, which were also available from the WordBank have been omitted due to different age ranges resulting in missing and outlier observations. Furthermore, observations beyond the 8-30 months’ age range were also omitted. The language acquisition forms allowed in the analysis are both words and gestures, and words and sentences.

There are in total 31,218 observations. The proportion of data from each language family is largely uneven: 23,016 observations from the Germanic languages, 4,518 observations form the Romance languages and 3,684 observations from the Slavic languages. However, this will not influence the results as the data does not follow a normal distribution (see below). There are in total 14,716 females, 15,027 males and 1,475 infants of unreported gender. The mean age of all the infants is 19.76 months, SD = 6.08. For the Germanic learning children, the mean age is 20.08, for the Romance learning children 18.92 and for the Slavic learning children 18.82 months.

## Analysis

Because the independent variables “production score” and “comprehension score” are count variables, as all data points are non-negative integer numbers and represent the number of words produced or comprehended by the child, the data follow a Poisson distribution and not a normal distribution. A Q-Q plot substantiates this. However, as the deviance is by far larger than the degrees of freedom, meaning there is overdispersion, a Quasi-Poisson analysis has been conducted to account for this. The analysis was conducted with planned contrasts in accordance with the hypotheses: Slavic languages compared to Germanic and Romance, and Romance languages compared to Germanic. Prior to this, however, an inspection of the overall means of the comprehension score and production score was performed to assure that a difference between the language families was actually present. An inspection of the means of the comprehension and production scores was also performed for the ages eight and 30 months. Furthermore, the beta estimates were exponentiated in order to interpret the results.

# Results

By inspecting the overall mean of the comprehension score for each language family, Slavic seems to have the overall highest score and Romance the lowest (see figure 1). By inspecting the overall mean of the production score however, Germanic seems to have the overall highest score and Romance the lowest (see figure 2). This does not fully coincide with the proposed hypotheses, however there does seem to exist a difference in language acquisition scores between the language families, and the analysis was still conducted in accordance with the hypotheses.

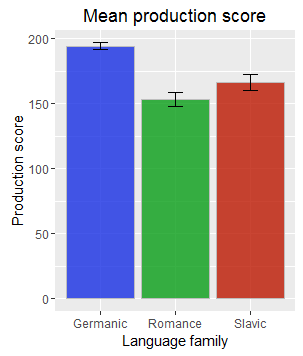
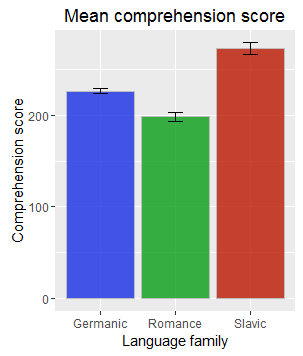


Figure 2

Figure 1

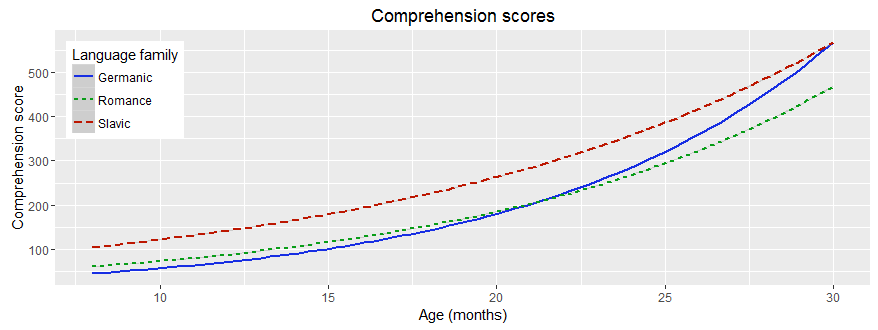
The model predicting language comprehension from the children’s age and their language family with planned contrasts significantly predicted the outcome: χ2(3) = 2,618,159, *p* < .001, R2 = .50 (Hosmer & Lemeshow). As would be expected, the age had a significant positive impact on the comprehension score: *b* = 0.10, *SE* = 0.0006, *t* = 172.16, *p* < .001. When comparing the Romance language family to the Germanic, the change had a significant positive impact on the comprehension score: *b* = 0.02, *SE* = 0.005, *t* = 4.12, *p* < .001. When comparing the Slavic language family to the Romance and Germanic language families, the change had a significant positive impact on the comprehension score: *b* = 0.10, *SE* = 0.003, *t* = 28.69, *p* < .001.

Figure 3

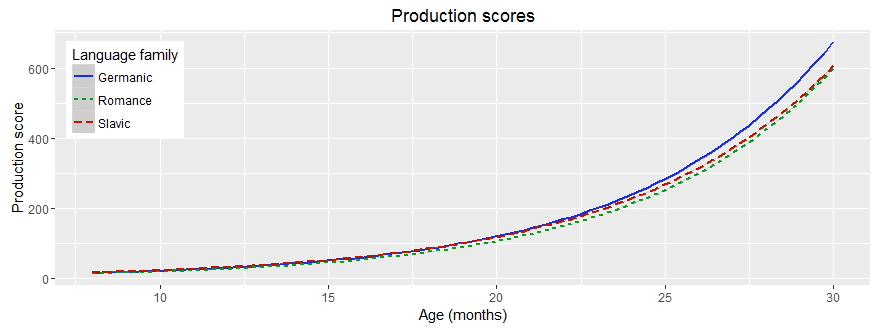
The model predicting language production from the children’s age and their language family with planned contrasts significantly predicted the outcome: χ2(3) = 4,791,851, *p* < .001, R2 = .66 (Hosmer & Lemeshow). Again, as would be expected, the age had a significant positive impact on the production score: *b* = 0.17, *SE* = 0.0008, *t* = 210.53, *p* < .001. When comparing the Romance language family to the Germanic, the change had a significant positive impact on the production score: *b* = 0.06, *SE* = 0.006,  *t* = 10.04, *p* < .001. However, when comparing the Slavic language family to the Romance and Germanic language families, the change had a highly non-significant negative impact on the production score: *b* = -0.0003, *SE* = 0.004, *t* = -0.06, *p* = .95.

Figure 4

Table 1  
The table displays the mean production and comprehension scores for each language family at eight and 30 months of age and the overall mean.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | **Mean (8 mo.)** | **Mean (30 mo.)** | **Overall mean** |
| **Germanic** | Production score | 3.01 | 509.92 | 194.27 |
| Comprehension score | 24.10 | 509.92 | 226.44 |
| **Romance** | Production score | 1.72 | 450.19 | 153.34 |
| Comprehension score | 41.85 | 450.19 | 198.48 |
| **Slavic** | Production score | 2.26 | 440.32 | 166.57 |
| Comprehension score | 43.88 | 480.75 | 273.00 |

# Discussion

The results suggest that there is a difference in language acquisition between language families. However, the difference is not as great in language production as in language comprehension and does not fully coincide with the proposed hypotheses.

Moving from the Germanic and Romance language families to the Slavic language family results in almost no difference at all in production score (-0.01 %, non-significant), but a 10 % increase in comprehension score. This partly supports the one hypothesis, as Slavic learning children seem to comprehend language faster than Romance and Germanic learning children, but there does not seem to be any real difference in production between the language families. Moving from the Romance to the Germanic language family results in a 6 % increase in production score and a 2 % increase in comprehension score. This does not support the second hypothesis, as Germanic learning children seem to acquire language faster then Romance learning children.

In line of this, the first hypothesis stating Slavic learning children will be faster than Germanic and Romance learning children is only partly supported, and the second hypothesis stating Romance learning children will be faster than Germanic learning children is not supported by the results.

The hypotheses, however, cannot fully be assessed by the models expressing the overall mean. To fully assess the hypotheses, the mean language acquisition scores at the beginning and end of the age range, eight and 30 months, respectively, must be inspected. By doing so, it will be clear whether one or some of the language families start out with a higher score, meaning the children from that language family have started acquiring language earlier or quicker than the children of the other language families. Furthermore, it will be clear whether one or some of the language families end with a higher score, meaning the children from that language family comprehend and/or produce more language and have reached a higher stage of language acquisition than the children from the other language families at the age of 30 months. This can be clarified by looking at table 1. As can be seen from table 1, Germanic learning children have the lowest mean comprehension score at the age of eight months and Slavic the highest, however at the age of 30 months, Germanic learning children have the highest mean comprehension and production score and Romance the lowest. At the age of eight months, the mean production score is almost the same for the three language families, however Germanic has the highest.

Taking all this into account suggests that Slavic learning children overall comprehend language faster than Germanic and Romance learning children, however, by the age of 30 months, Germanic learning children have reached a higher state of language comprehension than Romance and Slavic learning children. It is important to note, however, that Germanic learning children start out with the lowest comprehension score at the age of eight months. This suggests that Germanic learning children are slower at comprehending language at first, but perhaps better at “bootstrapping” and thus comprehending more words faster than Romance and Slavic learning children after a certain point. It seems to suggest that Germanic learning children have a greater vocabulary explosion.

The results also suggest that Romance learning children overall do not acquire language faster than Germanic learning children, neither in production nor comprehension, even though Romance learning children also start out with a higher comprehension score than Germanic learning children at the age of eight months. Overall, Romance learning children actually performed the worst of the three language families, thus the second hypothesis is not supported by this data.

The reason why Slavic learning children comprehend more words than children from the other language families both overall and at the age of eight months might be that most of the Slavic languages have word-initial fixed stress. This word segmentation cue is a quite strong cue as it has none to few exceptions, making it easier for infants to segment speech. This could give Slavic learning children a head start compared to the other languages, where the word stress cue is not as accurate. Furthermore, as Slavic languages (except Bulgarian and Macedonian) do not make use of articles and Dale & Goodman (2005) suggest that grammatical knowledge has influence on language acquisition, this might also be a factor as to why Slavic learning children are faster at comprehending language.

It is interesting that Germanic learning children acquire language faster than Romance learning children (except for comprehension at the age of eight months), when previous research indicate otherwise (e.g. Lleó & Demuth, 2015; Bleses, et al., 2008). As mentioned previously, Germanic learning children might be better at cracking the code for language acquisition, even though it happens at a later state than Romance and Slavic learning children. However, to figure this out, more research must be conducted on this area.

As mentioned previously, there is not much difference in language production between Slavic learning children and Germanic and Romance learning children. However, by looking at figure 3, this seems to be because Romance learning children perform worse than Slavic learning children, and Germanic learning children perform better, meaning the two language families even the score. To test whether this is true, an analysis comparing each language family to each other should be conducted.

A factor to consider when inspecting the results is that there is great variability in when children start to speak (Harley, 2010). Although there is a general trend showing that there is a difference between language families, the individual variability between children might affect the results, especially because each observation comes from a different child, so one cannot follow one child from beginning to end of the age range. On the other hand, with more than 30,000 observations, individual variability should not be an issue.

Also, a source of error to consider is that of the three Romance languages used in this analysis, Quebec French was one of them. French is in some ways different from the other Romance languages as the case of no null-subject clauses in French and influence from German in the French language. Furthermore, Quebec French might differ even more from the other Romance languages, as the language has evolved in a completely different region of the world (i.e. in North America) than the other Romance languages. For instance, Quebec French has more phonemes than Parisian French, which is the language spoken in France. (Quebec French Phonology, n.d.)

A limitation to this analysis is that there are only three languages in each language family, as no more data accessible from WordBank fitted the criteria. The small amount of languages in each language family might not reflect the true relationship between the language families because the language families are not fully represented. Furthermore, by using WordBank’s CDI data, the data relies fully on the parents’ ability to correctly observe and score their child’s language development. This is probably not the most accurate measure, however it is the most natural measure compared to observing a child in a laboratory by a professional. Here, ecological validity is of more importance than complete accuracy, and the fact that children observed in a laboratory or by a stranger might not act the way they would usually do, had they been in an environment they consider safe, this is probably the best option.

For further research, it would be interesting to include language families from all around the world in the analysis, and not just from the Indo-European language family. A comparison of the greater language families such as Indo-European versus Afro-Asiatic or Sino-Tibetan would also be interesting as to test whether there is a difference in language acquisition all around the world and not just from language family to language family within a certain area (as what have been done here). One might also include other factors in the comparison of language families, such as the gender of the child or the parents’ education to test whether there is a difference depending on gender or the parents’ education between the language families. Furthermore, comparing each language family to each other, instead of using planned contrasts, over time (i.e. an interaction effect model) might reveal additional knowledge to what has been found here.

# References

Bleses, D., Vach, W., Slott, M., Wehberg, S., Thomsen, P., Madsen, T. O., & Basbøll, H. (2008). Early vocabulary development in Danish and other languages: A CDI-based comparison. *Journal of Child Language*, 35, 619-650.

Cutler, A. & Carter, D. M. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech and Language*, 2, 133-142.

Dale, P., & Goodman, J. (2005). Commonality and individual differences in vocabulary growth. In M. Tomasello & D. I. Slobin (eds). *Beyond nature-nurture: Essays in honor of Elizabeth Bates*, 41-80.

DeCasper, A. J. & Spence, M. J. (1986). Prenatal maternal speech influences newborn's perception of speech sounds. *Infant Behavior and Development*, 9, 133-150.

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (in press). Wordbank: An open repository for developmental vocabulary data. *Journal of Child Language*.

Gerken, L. A. (1996). Prosodic structure in young children's language production. *Language*, 72, 683-712.

*Germanic Languages*. (n.d.). Retrieved Dec 17, 2016, from Wikipedia: https://en.wikipedia.org/wiki/Germanic\_languages

Hanuliková, A. (2009). *Studia grammatica: Lexical segmentation in Slovak and German.* De Gruyter.

Harley, T. A. (2010). *Talking the Talk.* Psychology Press.

Johnson, E. K. & Jusczyk, P. W. (2001). Word Segmentation by 8-Month-Olds: When Speech Cues Count More Than Statistics. *Journal of Memory and Language*, 44, 548-567.

Jusczyk, P. W. (1999). How infants begin to extract words from speech. *Trends in Cognitive Science*, 323-328.

Lleó, C. & Demuth, K. (2015). Prosodic Constraints on the Emergence of Grammatical Morphemes: Crosslinguistic Evidence from Germanic and Romance Languages. *researchgate.net*.

*Quebec French Phonology*. (n.d.). Retrieved Dec 25, 2016, from Wikipedia: https://en.wikipedia.org/wiki/Quebec\_French\_phonology

*Romance Languages*. (n.d.). Retrieved Dec 17, 2016, from Wikipedia: https://en.wikipedia.org/wiki/Romance\_languages

Rudder, J. (n.d.). *Comparative Grammar of the Romance Languages*. Retrieved Dec 20, 2016, from Nativlang: http://www.nativlang.com/romance-languages/

*Slavic Languages*. (n.d.). Retrieved Dec 17, 2016, from Wikipedia: https://en.wikipedia.org/wiki/Slavic\_languages

Wagner, J. (n.d.). *European Languages*. Retrieved Dec 26, 2016, from ielanguages: http://ielanguages.com/eurolang.html

Werker, J. F. (1995). Exploring Developmental Changes in Cross-Language Speech Perception. In D. N. Osherson, L. R. Gleitman, & M. Liberman, *An Invitation to Cognitive Science: Language* (pp. 87-106). A Bradford Book.