Second-Language Acquisition and Morphological Irregularity in Simulations of Natural Language Evolution

AP Research

May 19, 2021

Word Count: 5065

Second-Language Acquisition and Morphological Irregularity in Simulations of Natural Language Evolution

In linguistics, morphological compositionality and morphological irregularity are terms that refer to the manner in which a word is inflected. Compositionality is the representation of complex meanings via a combination of representations of their parts (Szabó, 2020). For instance, in the English phrase, "three red dogs," each of the four components, which are "three," "red," "dog," and "-s," can be understood in isolation, with each contributing to the overall meaning of the phrase. Compositional structure is very prevalent in natural language because speakers can easily combine components of meaning into a larger phrase, even if the phrase itself is new. In contrast, irregular inflection involves representations that lack such a trivial relation to their constituent parts (Wu et al., 2019). As an example, the simple past conjugation of the English verb "to have" is "had." This cannot be interpreted simply as the verb stem with the "-ed" suffix, which marks the simple past with regular verbs. For example, the compositional verb "to walk" has the simple past form "walked." Irregular forms must be individually learned as compositional patterns do not apply, which means that irregular structures are harder to acquire than compositional structures for second-language (L2) speakers (Birdsong & Flege, 2001).

It is understood that languages adapt to the settings in which they are used, following patterns that may be reminiscent of those observed in biological evolution (Croft, 2002). Consequently, one may wonder how such adaptation presents itself when examining compositionality and irregularity. More specifically, one point of curiosity surrounds how the presence of second-language speakers affects the degree of irregularity in natural languages. This is difficult to examine through cross-linguistic analyses because of logistical problems, in addition to a lack of relevant data. Given that they do not have these issues, simulated languages are employed to study this relationship in this text.

Literature Review

Linguistic Adaptation

The eminence of linguistic adaptation research is rather new. In the past, culture and linguistic structure were not often correlated (Perkins, 1992), although efforts did exist during this time period (e.g., Sapir, 1912). In the modern era, there is significantly more interest in the interplay of various factors that influence language evolution (Perkins, 1992). This has led to an abundance of research in this field (e.g., Bentz & Winter, 2012; Croft, 2002; Nettle, 1998; Wray & Grace, 2007), especially with respect to social structure.

In this context, Gary Lupyan and Rick Dale published an influential study about linguistic adaptation in 2010. They performed a cross-linguistic analysis of many grammatical factors and found that the social context in which a language is spoken influences its grammar. In subsequent research, this idea was expanded to include stressors like ecology, communication technology, and genetic factors (Lupyan & Dale, 2016). Within the 2010 study, they examined grammatical structure in relation to the physical distribution and number of speakers of a language, as well as the quantity of interlinguistic contact. In total, over two thousand languages were studied. The results indicated that social factors had an impact on the structure of language, suggesting that language adapts so that it is more suitable for communication in the contexts where it is used (Lupyan & Dale, 2010). One proposed explanation for this trend involved mechanisms of language acquisition, with the authors suggesting that languages with more adult learners are less grammatically complex because features that are difficult to acquire are less likely to be passed on to subsequent generations. On the other hand, languages with higher grammatical complexity are more redundant, which means that they are easier for infants and young children to acquire. These notions are supported by existing research in the field of L2 acquisition.

Second-Language Acquisition

Birdsong and Flege (2001), for example, examined the way that second-language speakers of English generate irregular forms such as irregular verb conjugations and irregular plural forms. They used samples of native Spanish and Korean speakers for this purpose. Participants were provided with sentences missing a word and were then asked to choose the proper form of the absent word. Their results indicated that low-frequency words were overgeneralized. In other words, a compositional inflection paradigm was used when an irregular paradigm would have been correct (Matiini, 2016). In addition, they found that the native language of a speaker affected their performance; the Korean speakers performed better with irregular verbs than the Spanish speakers did (Birdsong & Flege, 2001). Considering that Korean verb phrases are morphologically richer than those in Spanish (Sells, 1995), this suggests that the speakers' native languages shaped their abilities in English. These results provide further evidence to support Lupyan and Dale's explanation for their findings, since there would then be evolutionary pressure applied by L2 speakers that have trouble with certain features of a language. Regardless of this, these findings cannot establish a causal relationship between L2 speakers and morphological shifts.

Knowledge Gaps

Even though research that discusses irregularity relative to L2 acquisition exists (e.g., Birdsong & Flege, 2001; Housen, 2002), it only examines this for individual speakers and their cognitive processes. There is much less research about the effect of L2 speakers on linguistic change. Some studies do exist (e.g., Bentz et al., 2015, Bentz & Winter, 2012) but these do not examine irregularity However, the findings of these studies indicate that the impact of L2 speakers is often quite significant. This perspective is not universal; there is evidence suggesting that other factors, such as population size, are more important in determining linguistic structure (Koplenig, 2019). Despite this, the significant linguistic changes that L2 speakers frequently accompany further the notion that irregularity-based research in L2 acquisition should be conducted.

The lack of research in this area is likely due to the difficulty of acquiring suitable and reliable data. It is difficult to account for external factors, especially given the scope of this study. For instance, languages with a higher total number of speakers are more likely to have larger numbers of L2 speakers as well (Lupyan & Dale, 2010). Since languages with large population sizes undergo different evolutionary trends from those with smaller population sizes (Bromham et al., 2015; Koplenig, 2019), it is difficult to know whether population size or the presence of L2 speakers is driving evolution. In addition, interlinguistic contact with languages outside the realm of study is difficult to avoid, posing further logistical difficulty. These factors have limited this aspect of the body of knowledge within linguistics. This gap is significant since its presence limits the validity of Lupyan and Dale’s proposed mechanism for linguistic change in their 2010 study. Consequently, the goal of this study is to determine the nature of the correlation between the prominence of L2 speakers and morphological irregularity in the absence of interfering influences. This is accomplished through the examination of simulated language.

Computational Approaches

Simulations of natural language development have become increasingly common in recent years (Mitkov, 2005) since this practice has many advantages over its more traditional counterparts. As an example, simulated languages can be employed to study language as an interconnected system because all relevant aspects of these languages are recorded or can be observed from the examination of samples (Cangelosi & Parisi, 2002). The same cannot be said for real languages, since the availability of appropriate language samples is often a limiting factor in linguistic research (Mannila et al., 2013). Likewise, simulated language samples can be generated in large supply, and can be tailored to the requirements of the study (Cangelosi & Parisi, 2002).

Simulated languages are also useful because they can be altered as much or as little as one desires. On one hand, they can arise from disorder without outside intervention (e.g., Kirby, 2001). Although a grammar system and some degree of internal logic on the part of the speakers is required to accomplish this, no intervention is required once the generation process commences. On the other hand, the researcher can alter their grammar at any point, something that is impossible to do with natural language. Moreover, speakers of any two simulated languages can interact, which enables the observation of sociolinguistic changes. For this study, the process of L2 acquisition is of great interest in this regard since known proportions of L2 speakers can be introduced in a controlled manner.

Research Method

Data are gleaned from monitoring the behaviour of an iterated learning model (ILM) that is employed to simulate the development of natural language. The languages produced by the ILM, which is based on that described by Kirby (2001) have two fundamental components: a meaning space and a signal space. The other features of the ILM revolve around these two elements, with grammars converting elements of the meaning space to those of the signal space. For the purposes of this study, data are collected by counting the number of irregular forms in various languages situated in different contexts. The goal is to allow the ILM to adapt to its environment in order to observe how these contextual shifts affect the degree of irregularity in the simulated languages. This method is aligned with the research question because these two factors can be observed in an environment free from other influences.

Arbitrariness

Although most of the behaviour of the ILM is designed to be as organic as possible, there are still some parameters that are arbitrarily defined. These are included in Appendix A. Ideally, the arbitrariness of parameters of a language evolution model should be as minimal as possible (Cangelosi & Parisi, 2002). However, these values must be held constant in order to attain uniform, meaningful data. In addition, portions of the simulation were executed with reasonable variation in all parameters and consistent results were obtained. This was also observed by Kirby (2001) regarding the parameters of their ILM. These findings indicate that results should allude to fundamental patterns in the languages themselves and that patterns are not simply dependent upon these fixed parameters.

Meaning Space

The meaning space for the ILM is two-dimensional with 10 possible values for each of the two components, termed *a* and *b*. Therefore, there are 100 possible meanings, ranging from *(a_0, b_0)* to *(a_9, b_9)*. This is different from Kirby's approach, as their ILM uses a meaning space with five possible values for each component. The larger meaning space is used in order to provide more graduation in terms of the number of compositional meanings in a language, while still ensuring that the speed of the simulation does not suffer too heavily. In terms of real-world parallels, the elements of the meaning space can be thought of as verb-object pairs indicating simple phrases or verbal inflections for different grammatical person and number (Kirby, 2001).

Signal Space

The signal space contains sequences of lowercase letters. Signals are linear and ordered. That is, letters are verbalized one at a time and changing the order of letters can change the expressed meaning. Perhaps counterintuitively, each individual character does not necessarily represent a letter or a corresponding sound. As stated by Kirby (2001), they are the "atomic elements of the language that the grammatical system cannot break down." (p. 103)

Internal Representation of Grammar

Each agent in the ILM has an individual representation of grammar. This enables the agent to convert from the meaning space to corresponding signals and allows for the generation of new signals. Although the specifics of this system are not required to appreciate its overall behaviour, it is fruitful to discuss its two most important algorithms, since these are responsible for ensuring that the agents behave similarly to human speakers. The two algorithms in question are the induction algorithm and the invention algorithm. For the sake of brevity and clarity, these procedures are not described in detail. A more comprehensive discussion is available in Appendix B.

The induction algorithm is responsible for generalizing rules in a compositional fashion. It lets agents identify patterns in input so that they can infer signals for meanings that they have not encountered, provided that they have encountered the components in isolation. This structure allows for the development of compositional grammars, which are vital if the simulated languages are to emulate natural language.

The invention algorithm dictates how agents produce new signals when their grammar cannot produce a signal for a certain meaning. Although random signals with lengths ranging from one to 10 characters are generated for this purpose, the algorithm does not always assign a random signal to the entire meaning. Rather, the grammar will first try to find compositional rules that correspond to one component of the meaning. From here, it will associate the random signal with the other component. If no such rule exists, then the random signal will represent the whole meaning. This process maintains the compositional rules derived by the induction algorithm.

It is also important to note that if an agent can form multiple signals for the same meaning through different means, it will always produce the shortest valid signal. This is to impose a pressure towards efficiency, which is present within the communicative niche of natural language (Gibson et al., 2019). Such an adjustment leads to more realistic approximations of natural language (Kirby, 2001).

Generation Structure

The evolution of language in the ILM is divided into generations. Although the specific structure of a generation varies, all generations are executed according to the same fundamental principles. In every generation, there are one or more speakers and one learner. The learner will receive a series of signals produced by the speakers. It is assumed that all agents involved in a transaction are aware of the meanings represented by the produced signal due to some external knowledge or a stimulus that both parties are experiencing. This means that the listener can induce the meaning-signal pairs into its personal grammar. After all transmissions have occurred, the learner becomes a speaker for the next generation, in which the same process, or a variation thereof, occurs.

The elements of the meaning space that are used in conversation are distributed unevenly. They occur according to Zipf's law, which states that the frequency of a given word is approximately proportional to the reciprocal of its frequency rank (Piantadosi, 2014). In other words, the second-most common word would be half as frequent as the most common word, and the third-most common word would be one-third as frequent as the most common word. For the purposes of the ILM, this pattern is defined based on the values of i and j for a meaning (a_i, b_j), with meanings with smaller values of i and j more likely to occur. Since many common words do not have an independent meaning and serve syntactic functions instead, Zipf's law is not always applicable in the realm of natural language (Piantadosi, 2014). However, the distributions of many natural phenomena obey Zipf's law (Corominas-Murtra & Solé, 2010). In addition, enacting a Zipfian distribution over the meaning space leads to closer approximations of natural language because it makes compositionality more beneficial (Kirby, 2001). In such a scenario, it is unlikely that an agent would encounter meanings with high values of *i* and *j*, which means it must be able to infer signals. Therefore, a Zipfian distribution is beneficial for this simulation.

Another quirk of the ILM involves the replication of real-life speaker error and dialectal shifts. There is a small probability that any given character in a signal is not articulated by a speaker. In this case, the learner induces the eroded signal into its grammar, which helps lead to the development of irregularity (Kirby, 2001).

Simulation Structure

The simulation process is divided into two phases: the homogeneous phase, in which the initial evolution of language occurs; and the heterogeneous phase, in which interlinguistic contact materializes. For all runs of each phase, 100 generations are analyzed and the number of irregular forms per generation is recorded. In order for a generation to be included in analyses, the listener has to be able to produce signals for all meanings without inventing new signals, which ensures consistency throughout the data.

Every simulation is performed with a seed from which all necessary random values are generated to enable the repetition of specific simulations if desired or if necessary. The seed that generated the data for this study is found in Appendix A.

Homogeneous Phase

The homogeneous phase is responsible for the establishment of grammars that are later employed for the examination of L2 acquisition in the heterogeneous phase. More specifically, languages are evolved until a certain proportion of their vocabulary matches that of an earlier version of the language to ensure that languages have reached a baseline degree of stability. It is important to note that this metric takes the relative frequency of meanings into consideration. Thus, for intelligibility purposes, having similarities among common meanings is better than having similarities among rare meanings.

The homogeneous phase is run with 10 random seeds generated from the simulation seed and thus generates 10 stable grammars.

Heterogeneous Phase

In the heterogeneous phase, speakers of two languages are combined in order to analyze changes in the prevalence of irregularity due to L2 acquisition. Speakers of the least irregular language from the homogeneous phase learn the most irregular language, and vice versa. For this purpose, irregularity is quantified by taking the mean of the number of irregular forms across all samples of a language. Although it would be ideal to model communication between speakers of all possible combinations of languages, this is not feasible due to the speed at which this process proceeds. For each of the native languages, L2 speakers are introduced in 5% intervals, which results in 21 runs for the acquisition of each language, with the proportion of L2 speakers ranging from 0% to 100%. This introduction does not occur over time for the same language sample; each of the 21 populations for each native language are wholly distinct. The graduation of the L2 speaker proportion is important because it enables a more detailed examination of how L2 speakers affect linguistic development depending on the prominence of their presence.

Results

Throughout the progression of the ILM, it was evident that the produced grammars were undergoing adaptation. This is demonstrated by examining the signals produced by a sample grammar from a stable language.

Table 1

Signals from a Sample Language

a component *b* component	a_0	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9
b_0	*x*	*ts*	*u*	*e*	*to*	*t*	*n*	*tf*	*ti*	*tna*
b_1	*px*	*ps*	*p*	*pe*	*po*	*pp*	*pn*	*pf*	*pi*	*pna*
b_2	*h*	*hs*	*hu*	*he*	*ho*	*hp*	*hn*	*hf*	*hi*	*hna*
b_3	*xx*	*xs*	*xu*	*xe*	*xo*	*xp*	*xn*	*xf*	*xi*	*xna*
b_4	*c*	*cs*	*cu*	*ce*	*co*	*cp*	*cn*	*cf*	*ci*	*cna*
b_5	*sx*	*ss*	*su*	*se*	*so*	*sp*	*sn*	*sf*	*si*	*sna*
b_6	*bx*	*bs*	*bu*	*be*	*bo*	*bp*	*bn*	*bf*	*bi*	*bna*
b_7	*a*	*as*	*au*	*ae*	*ao*	*ap*	*an*	*af*	*ai*	*ana*
b_8	*nx*	*ns*	*nu*	*ne*	*no*	*np*	*nn*	*nf*	*ni*	*nna*
b_9	*ix*	*is*	*iu*	*ie*	*io*	*ip*	*in*	*if*	*ii*	*ina*

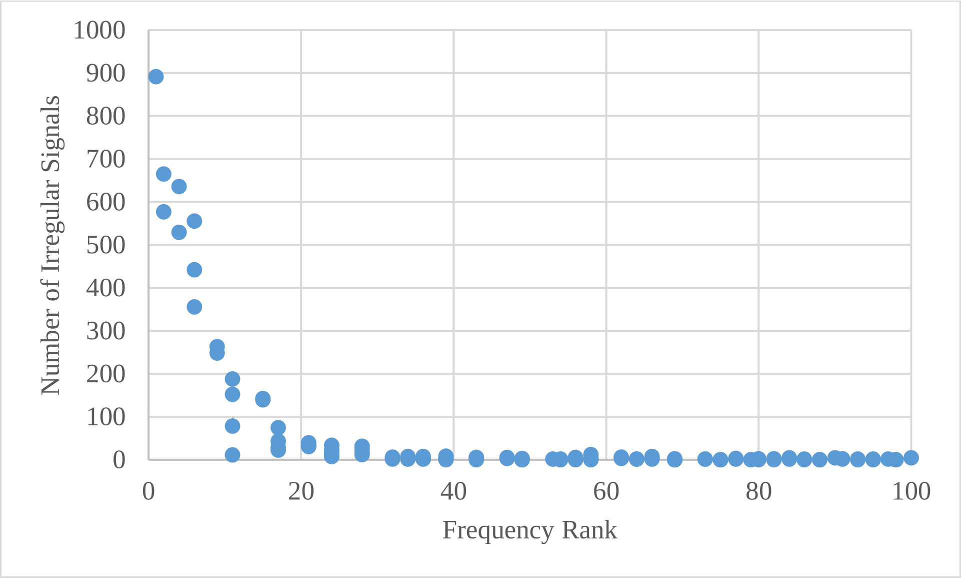
All of the signals are quite short, in spite of the fact that randomly generated signals could be up to 10 characters long. This is due to the preference of shorter signals by the ILM. In this respect, the ILM is acting in accordance with Lupyan and Dale's findings because the grammar has adapted to the situation; shorter signals are transmitted more readily so they become more prevalent. The fact that this adaptive tendency is present without external intervention may suggest that the real-world parallels of other results are more likely to be valid.

Homogeneous Phase Results

In the homogeneous phase languages, the frequency of irregularity was found to be higher in more common meanings, as demonstrated by the following graph.

Figure 2

Comparing Frequency and Irregularity



The x-axis corresponds to the frequency rank of a meaning, whereas the y-axis depicts the number of irregular signals for that meaning. This highlights the correlation between frequency and the likelihood of irregularity, which is evidenced by the drastic decline in the number of irregular signals as frequency rank increases. Since there were 10 languages in the homogeneous phase, each with 100 collected samples, there were 1000 samples collected in total. Therefore, (a_0, b_0), the most frequently produced meaning, was irregular in almost 90 percent of samples. This result is important because it parallels the correlation between frequency and irregularity that is observed in natural language (Wu et al., 2019). In spite of this, no pressure was placed on the ILM for this to be the case (Kirby, 2001). This, in conjunction with the relative lengths of signals as previously discussed, supports the validity of the application of these results to natural language.

Across all samples from the homogeneous phase, the mean number of irregular forms was 6.44 ± 0.09 with 95% confidence and the median number of irregular forms was 7. The irregularity of individual languages was more varied than that of the amalgamation of samples from all of the languages, which is highlighted in the following table.

Table 3

Irregularity Data by Language

Language number	Mean number of irregular forms (95% confidence)	Median number of irregular forms
0	7.25 ± 0.27	7
1	6.48 ± 0.18	6
2	7.0 ± 0.15	7
3	7.07 ± 0.22	7
4	7.42 ± 0.23	7
5	6.84 ± 0.21	7
6	5.6 ± 0.21	6
7	4.72 ± 0.33	4
8	5.64 ± 0.29	6
9	6.38 ± 0.22	7

The number of irregular forms was plotted as a function of the number of generations after stability was reached in order to determine if these two values were correlated. This demonstrated that there was no significant relationship between these two metrics; there were periodic fluctuations, but these did not seem to be absolute in any case.

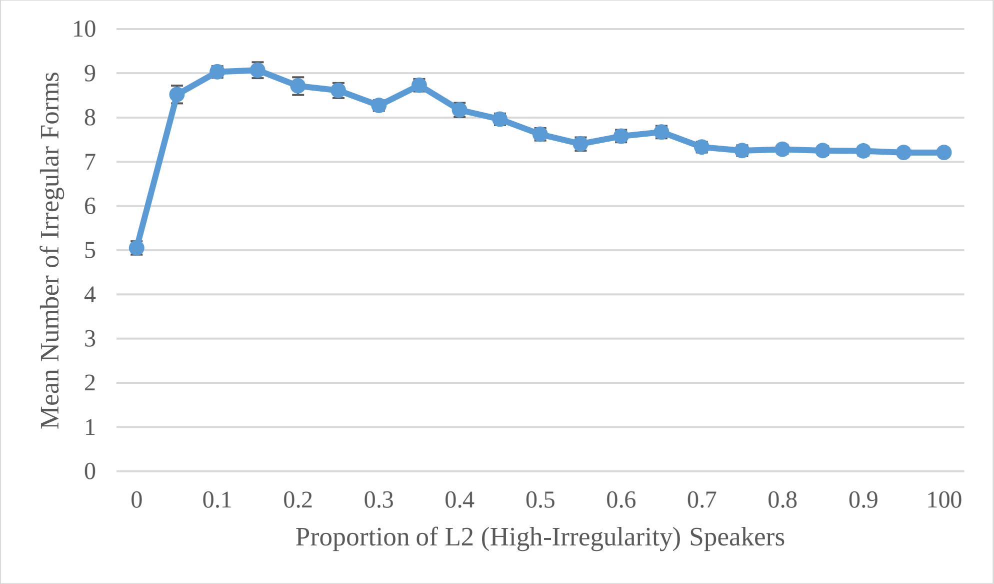
Heterogeneous Phase Results

The heterogeneous phase was executed with language 7 as the low-irregularity language and language 4 as the high-irregularity language. The discrepancy in irregularity between these two languages, as highlighted by the collected data, was large enough to highlight the dynamics of L2 acquisition and its effect on irregularity.

As expected, the introduction of second-language speakers affected the degree of irregularity in the language. However, the relationship between these two variables is not as simple as one might assume. This is illustrated by the graph below, which details the acquisition of the low-irregularity language by speakers of the high-irregularity language.

Figure 4

Acquisition of Low-Irregularity Language



This graph indicates that contrary to prior expectations, irregularity is not greatest when the highest proportion of high-irregularity language speakers is present. Rather, irregularity is most prominent with low, nonzero proportions of L2 speakers. This is likely because irregular forms are more readily transmitted cross-linguistically than compositional forms since irregular signals are often shorter than their compositional counterparts. To illustrate this, consider the following two hypothetical languages.

Figure 5

Signals Produced by Hypothetical Grammars

	a_0	a_1
b_0	*ab*	*x*
b_1	*eb*	*ed*

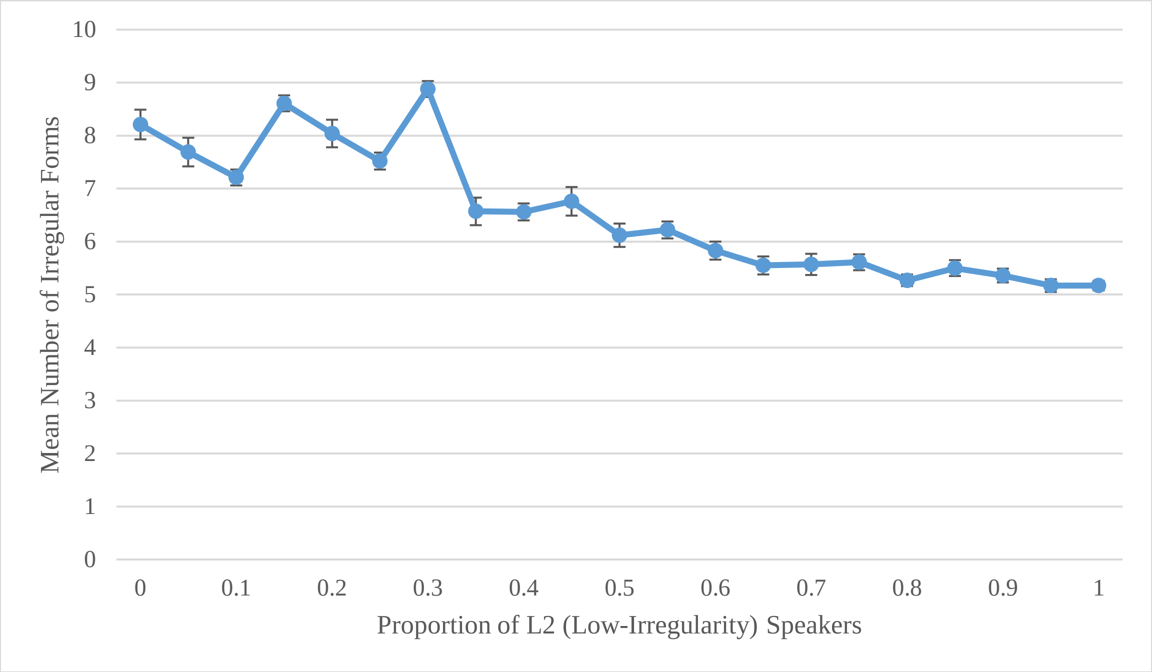
	a_0	a_1
b_0	*Ip*	*iq*
b_1	*y*	*oq*

Suppose that a learner has encountered all signals in each language. Since the ILM will produce the shortest possible signal for a meaning, the learner will signal *x* for the meaning (a_1,b_0) and *y* for the meaning (a_0, b_1). Therefore, the grammar formulated by the learner will be more irregular than either of the parent grammars. This phenomenon is the likely reason for the unexpected increase in irregularity, which is also intensified by the fact that the L2 speakers spoke a high-irregularity language, meaning that irregular signals were more common in the first place. As the proportion of L2 speakers increases further, learners are less likely to acquire irregular forms present in the first language since they are transmitted less frequently, which is why the degree of irregularity decreases.

With low but non-zero proportions of L2 speakers, the high-irregularity acquisition data demonstrate trends similar to those in the low-irregularity data.

Figure 6

Acquisition of High-Irregularity Language



As with the acquisition of the low-irregularity language, there are increases in irregularity with low proportions of L2 speakers. In this case, though, these are offset by the trend towards compositionality. This occurs as more low-irregularity speakers are introduced. In addition, the degree of irregularity of these samples was much more variable than those involving the acquisition of the low-irregularity language, which is evidenced by larger standard deviations and confidence intervals. This indicates that high-irregularity languages have more varied behaviour since the pressures of keeping the grammar constant and simplifying it for easier acquisition are in competition with each other.

The fact that such wide variations in the prevalence of irregularity were observed indicates that the languages were adapting to their circumstances and that such discrepancies are not simply a result of random chance. According to hypotheses of linguistic adaptation (e.g., Lupyan & Dale, 2016), the degree of irregularity in the languages is changing as a result of the proportion of L2 speakers changing. Via the same theoretical frameworks, this is justified through the notion that the changes in the degree of irregularity facilitate easier acquisition of the language for future generations.

It should be noted that these data are only representative of samples gleaned from the use of the same random seed. More specifically, if the same process of data collection was repeated for further generations produced from the same seed, the sample mean of the new data would lie within the confidence interval 95% of the time. This would not necessarily be the case if different seeds were employed; in such a scenario, sets of data could have different sample means and confidence intervals.

Discussion

Implications

The analyses of the data support the notion that the proportion of L2 speakers affects morphological irregularity. These two variables were observed in the absence of potential interferences like population size, geographic distribution, and other sociocultural factors beyond the different native languages. Thus, it is quite likely that the changes in the prominence of irregularity were due to the alterations in the proportion of L2 speakers. This may serve as an addition of information to a sparse area in the body of knowledge. As well, the data support the mechanism for the emergence of compositionality proposed by Lupyan and Dale. The data also support the conclusion drawn by Birdsong and Flege regarding L2 acquisition in their 2001 study; that is, the structure of native languages appears to influence how additional languages are interpreted by speakers.

In a broader linguistic context, the behaviour of the ILM supports an existing mechanism for suppletion, which occurs when different forms of a word have unrelated etymologies (Mel'čuk, 2006). In the ILM, suppletion occurred rather prominently; it took place when speakers of two languages induced meanings from both languages, not just their own. This parallels the importance of linguistic borrowing for suppletion in natural language (Maiden, 2004). Despite the high frequency of suppletion in natural language, it is often left unanalyzed (Bobaljik, 2014). Moreover, contemporary research has expressed the potential utility of suppletion-based analyses for cognitive representations of grammar (Bobaljik, 2014). This, coupled with the important role of suppletion in dictating irregularity within the simulated languages, indicates that suppletion may be more important than previously thought.

Although this study involved simulated languages, the behaviour of the ILM indicates that these results are likely reflected within natural languages. In many respects, the behaviour of the model corresponds to that of natural languages without explicitly being pressured to do so. Some notable examples of this are: (a) the generation of language from randomness, which mirrors how humans were able to spontaneously develop language; (b) the increased frequency of irregularity in common meanings, which supports conclusions made based on cross-linguistic data (Wu et al., 2019); and (c) the adaptation of simulated grammars to factors like L2 speakers and a pressure to produce shorter signals, which occurs in accordance with notions of linguistic adaptation (Lupyan & Dale, 2016). Since these tendencies were not arbitrarily imposed onto the model, it is perhaps the case that trends such as those observed with L2 speakers and morphological irregularity are also present in the real world. However, the model cannot produce natural languages, which means that this is not guaranteed. This inability is a major limitation of the study.

Limitations

The data collected throughout the course of this research are fundamentally limited because they were gleaned from simulated languages rather than natural languages. Simulations of natural language, no matter how elaborate, are incapable of replicating the countless nuances observed in natural languages (Cangelosi & Parisi, 2002). One particularly important assumption is that meanings are identical across groups of speakers. Linguistic evidence has demonstrated that this does not universally occur with respect to natural language (Thompson et al., 2018). Similarly, inevitable arbitrariness limits the degree of nuance that can be developed in the simulated languages (Cangelosi & Parisi, 2002), even though this did not appear to affect the nature of the results.

Another potential cause for concern arises when one considers that a relatively small meaning space was employed; the 100-element meaning space used by the ILM pales in comparison to the vast number of meanings expressed via natural language (Gong & Shuai, 2013). In a similar vein, due to computational restrictions, the number of scenarios for which data could be collected was limited. As an example, for each proportion of L2 speakers in each of the acquired languages, only one run of the simulation was performed. This limited the degree to which detailed quantitative analysis could be employed.

The social interactions present in the ILM were also rather rudimentary. Exchanges between agents from the same generation did not occur; language evolution was driven by intergenerational interactions. This made the correlation of L2 speaker prominence and irregularity easier to identify. Conversely, it inhibits the extension of the findings to natural language. This is because a more dynamic social structure would have made the setting in which the languages were spoken significantly closer to that present in the real world (Milroy & Milroy, 1997), especially since communication within a generation has important ramifications for L2 speakers (Bayram et al., 2019).

Future Directions

The findings of this study indicate that the role of L2 speakers in dictating the morphological irregularity of a language is substantial. Thus, the examination of this relationship as it occurs in natural language suggests itself as a possible direction for future research. Cross-linguistic analyses were beyond the scope of this study but they are a feasible vehicle for attaining knowledge, provided that sufficient data is available (Mannila et al., 2013). Moreover, historical analyses of suppletion could be conducted in a similar manner. These could be beneficial, given the importance of suppletion in dictating irregularity within the simulated languages.

Even within simulations of natural language, there is room for expansion. Computational models that employ more complex meaning spaces have been developed (e.g., Kirby et al., 2004). These meaning spaces, although still approximations of meaning in natural language, can facilitate more complex behaviour than that of the ILM developed for this study. Similarly, signals could be allowed to take on multiple meanings in different grammatical categories depending on context. This is common in natural language, with examples in English including "need," "mine," and "supply." The presence of such signals could influence the development of irregularity as L2 speakers would have to be able to discern which meaning the signal conveys, which has been demonstrated to be difficult (Jacobson et al., 2011). This could lead to stronger tendencies towards either compositionality or irregularity, which would make the presence of these signals valuable for the purposes of examining irregularity through the lens of L2 acquisition.

Conclusion

According to the findings of this study, the influence of L2 speakers on the balance between morphological irregularity and compositionality appears to be significant. This is situated in conjunction with previous research stating that the presence of L2 speakers affects other aspects of grammatical structure. It is evident that the investigation of natural languages is required in order to further examine this correlation. However, the behaviour of these two factors in isolation from potential interferences and the natural behaviour of the ILM in other regards suggest that L2 acquisition may have similar effects on morphological irregularity in natural languages.

References

Bayram, F., Pascual y Cabo, D., & Rothman, J. (2019) Intra-generational attrition: Contributions to heritage speaker competence. In B. Köpke & M. S. Schmid (Eds.), *Oxford handbook of language attrition* (pp. 446–457). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198793595.013.35

Bentz, C., Verkerk, A., Kiela, D., Hill, F., Buttery, P. (2015). Adaptive communication: Languages with more non-native speakers tend to have fewer word forms. *PLOS ONE, 10*(6). https://doi.org/10.1371/journal.pone.0128254

Bentz, C., & Winter, B. (2012). The impact of L2 speakers on the evolution of case marking. In T. C. Scott-Phillips, M. Tamariz, E. A. Cartmill, & J. R. Hurford (Eds.), *The evolution of language* (pp. 58–63). World Scientific Publishing. https://doi.org/10.1142/9789814401500_0008

Bobaljik, J. D. (2014). Suppletion: Some theoretical implications. *Annual Review of Linguistics, 1(*1), 1–18. https://doi.org/10.1146/annurev-linguist-030514-125157

Birdsong, D., & Flege, J. E. (2001). Regular-irregular dissociations in L2 acquisition of English morphology. *BUCLD 25: Proceedings of the 25th Annual Boston University Conference on Language Development* (pp. 123–132). Cascadilla Press.

Bromham, L., Hua, X., Fitzpatrick, T. G., & Greenhill, S. J. (2015). Rate of language evolution is affected by population size. *PNAS, 112*(7), 2097–2102. https://doi.org/10.1073/pnas.1419704112

Cangelosi A., & Parisi D. (2002). Computer simulation: A new scientific approach to the study of language evolution. In Cangelosi A. & Parisi D. (Eds.), *Simulating the evolution of language* (pp. 3–28). Springer. https://doi.org/10.1007/978-1-4471-0663-0_1

Corominas-Murtra, B., & Solé, R. V. (2010). Universality of Zipf's law. *Physical Review E, 82*(1), https://doi.org/10.1103/PhysRevE.82.011102

Croft, W. (2002). The Darwinization of linguistics. *Selection, 3*(1), 75–91. https://doi.org/10.1556/select.3.2002.1.7

Gibson, E., Futrell, R., Piantadosi, S. P., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How efficiency shapes human language. *Trends in Cognitive Sciences, 23*(5), 389–407. https://doi.org/10.1016/j.tics.2019.02.003

Gong, T., & Shuai, L. (2013). Computer simulation as a scientific approach in evolutionary linguistics. *Language Sciences, 40*(1), 12–23. https://doi.org/10.1016/j.langsci.2013.04.002

Housen, A. (2002). Verb semantics and the acquisition of tense-aspect in L2 English. *Studia Linguistica*, *54*(2), 249–259. https://doi.org/10.1111/1467-9582.00064

Jacobson, J., Lapp, D., & Flood, J. A seven-step instructional plan for teaching English-language learners to comprehend and use homonyms, homophones, and homographs. *Journal of Adolescent & Adult Literacy, 51*(2), 98–111. https://doi.org/10.1598/JAAL.51.2.2

Kirby, S. (2001). Spontaneous evolution of linguistic structure—An iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation*, *5*(2), 102–110. https://doi.org/10.1109/4235.918430

Kirby, S. (2002). Learning, bottlenecks and the evolution of recursive syntax. In T. Briscoe (Ed.), *Linguistic evolution through language acquisition* (pp. 173–204). Cambridge University Press. https://doi.org/10.1017/CBO9780511486524.006

Kirby, S., Smith, K., & Brighton, H. (2004). From UG to universals: Linguistic adaptation through iterated learning. *Studies in Language*, *28*(3), 587–607. https://doi.org/10.1075/sl.28.3.09kir

Koplenig, A. (2019). Language structure is influenced by the number of speakers but seemingly not by the proportion of non-native speakers. *Royal Society Open Science*, *6*(2). https://doi.org/10.1098/rsos.181274

Lupyan, G., & Dale, R. (2010). Language structure is partly determined by social structure. *PLOS ONE*, *5*(1). https://doi.org/10.1371/journal.pone.0008559

Lupyan, G., & Dale, R. (2016). Why are there different languages? The role of adaptation in linguistic diversity. *Trends in Cognitive Sciences*, *20*(9), 649–660. https://doi.org/10.1016/j.tics.2016.07.005

Maiden, M. (2004). When lexemes become allomorphs - On the genesis of suppletion. *Folia Linguistica*, *38*(3–4), 227–256. https://doi.org/10.1515/flin.2004.38.3-4.227

Mannila, H., Nevalainen, T., & Raumolin-Brunberg, H. (2013). Quantifying variation and estimating the effects of sample size on the frequencies of linguistic variables. In M. Krug & J. Schlüter (Eds.), *Research methods in language variation and change* (pp. 337–360). Cambridge University Press.

Matiini, G. (2016). Overgeneralization in singular/plural nouns and suffixed nouns of IELTS course students. *Journal of Language and Literature Education, 16*(2), 144–159. https://doi.org/10.17509/bs_jpbsp.v16i2.4478

Mel'čuk, I. (2006). Suppletion. In D. Beck (Ed.), *Aspects of the Theory of Morphology* (pp. 405–468). De Gruyter Mouton. https://doi.org/10.1515/9783110199864.2.405

Milroy, J, & Milroy, L. (1997). Network structure and linguistic change. In N. Coupland & A. Jaworski (Eds.), *Sociolinguistics* (pp. 199–211). Macmillan Publishers. https://doi.org/10.1007/978-1-349-25582-5_17

Mitkov, R. (Ed.). (2005). *The Oxford handbook of computational linguistics*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199276349.001.0001

Nettle, D. (1998) Coevolution of phonology and the lexicon in twelve languages of West Africa. *Journal of Quantitative Linguistics, 5*(3), 240–245. https://doi.org/10.1080/09296179808590132

Perkins, R. D. (1992). *Deixis, grammar, and culture.* John Benjamins Publishing Company.

Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review, 21*(5), 1112–1130. https://doi.org/10.3758/s13423-014-0585-6

Sapir, E. (1912). Language and environment. *American Anthropologist, 14*(2), 226–242.

Sells, P. (1995). Korean and Japanese morphology from a lexical perspective. *Linguistic Inquiry, 26*(2), 277–325.

Szabó, Z. G. (2020). Compositionality. *The Stanford encyclopedia of philosophy,* E. N. Zalta (Ed.).

Thompson, B., Roberts, S. G., & Lupyan, G. (2018). Quantifying semantic similarity across languages. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society : CogSci 2018* (pp. 2554–2559). Cognitive Science Society. https://cognitivesciencesociety.org/past-conferences/

Wray, A., & Grace, G. W. (2007). The consequences of talking to strangers: Evolutionary corollaries of socio-cultural influences on linguistic form. *Lingua, 117*(3), 543–578. https://doi.org/10.1016/j.lingua.2005.05.005

Wu, S., Cotterell, R., and O'Donnell, T.J. (2019). Morphological irregularity correlates with frequency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5117–5126. Florence, Italy.

Appendix A

Execution Log

This appendix contains the execution log for the program. Within the log, the simulation seed, as well as the seeds for individual runs are present.

File path: data/log.txt

File contents:

```
Execution Log
Seed for this execution: 1095409035377586249

Simulation seeds:
Homogeneous runs:
Simulation 0 seed: 5975092213289034014
Simulation 1 seed: -6630800722976653779
Simulation 2 seed: -2689288117522792040
Simulation 3 seed: 3635283437131530701
Simulation 4 seed: 3960364686693770430
Simulation 5 seed: -4188887271953292292
Simulation 6 seed: 1919544952033194253
Simulation 7 seed: -7970054014651408778
Simulation 8 seed: -4310711826927268857
Simulation 9 seed: 4613133846998227305
Language 7 was the least irregular and language 4 was the most irregular

Heterogeneous runs (high to low):
Simulation 0 seed: -1700109271936933871, L1 agent seed: 2827671166311071216, L2 agent
seed: 88150970101619869
Simulation 5 seed: -1547055657010746151, L1 agent seed: -1172588865290480501, L2
agent seed: 8753481618365160878
Simulation 10 seed: -1106720541297140020, L1 agent seed: -2585688078611322381, L2
agent seed: -3808893080084718264
Simulation 15 seed: -5215802370697000789, L1 agent seed: 1202436181104463362, L2
agent seed: 1634454123912511009
Simulation 20 seed: -2925304743566177672, L1 agent seed: -7096283374598077468, L2
agent seed: -2077275680175974065
Simulation 25 seed: -8160606667222815470, L1 agent seed: -344752210406803921, L2
agent seed: -7048311608039388455
Simulation 30 seed: -5928959691482860564, L1 agent seed: -8270414856626498245, L2
agent seed: -1130999154272822043
Simulation 35 seed: -7777165162140738544, L1 agent seed: -8783188963890367055, L2
agent seed: 2324356683370447953
Simulation 40 seed: 8224182336265143799, L1 agent seed: -192848303082515054, L2 agent
seed: -3918415127914640484
```

```
Simulation 45 seed: 7863001466262133712, L1 agent seed: 3876308308044610596, L2 agent
seed: 8900879229853901165
Simulation 50 seed: 5875789045212166225, L1 agent seed: 4207634780656862583, L2 agent
seed: -5342432210582628848
Simulation 55 seed: -1745557010478321104, L1 agent seed: 1315197458438497429, L2
agent seed: -658166239320279094
Simulation 60 seed: -6036006720972403004, L1 agent seed: -3656477991319859342, L2
agent seed: 6163956513097531346
Simulation 65 seed: -2957771626620958065, L1 agent seed: -8408368527085113812, L2
agent seed: -8820953784885644725
Simulation 70 seed: 7789718236071391383, L1 agent seed: 6883853231739804497, L2 agent
seed: 7417984624865804696
Simulation 75 seed: -8964561399495495208, L1 agent seed: -8839545399074368235, L2
agent seed: -2702992771261021728
Simulation 80 seed: -8858924799805119428, L1 agent seed: -7900607348125548612, L2
agent seed: 6328859994673451939
Simulation 85 seed: -4727336119528074033, L1 agent seed: 4387680442768349103, L2
agent seed: -1180983768512564070
Simulation 90 seed: 7522838258230606650, L1 agent seed: 583278443534466549, L2 agent
seed: -2073765880791995280
Simulation 95 seed: 3592032613278772380, L1 agent seed: 592899847511258105, L2 agent
seed: -2152237514682671379
Simulation 100 seed: 4687177426767936092, L1 agent seed: -2610423145986771985, L2
agent seed: -7592479862659268702
Heterogeneous runs (low to high):
Simulation 0 seed: -4952124015578856042, L1 agent seed: 8760241484134148409, L2 agent
seed: 6332917415982564164
Simulation 5 seed: 3290958445263522331, L1 agent seed: 1917148685882486931, L2 agent
seed: 6015619064091248943
Simulation 10 seed: -2855133701382681519, L1 agent seed: -6607831155906852369, L2
agent seed: 1505247171472237515
Simulation 15 seed: -8009457811332875446, L1 agent seed: 4655478977582257742, L2
agent seed: -4899903446581219125
Simulation 20 seed: 4916033694462168298, L1 agent seed: 6927893553259096569, L2 agent
seed: -3542216735937303943
Simulation 25 seed: 7137999422942342145, L1 agent seed: -8634510879121082401, L2
agent seed: -558983996851547758
Simulation 30 seed: -5329560313167277934, L1 agent seed: 3448841627017508104, L2
agent seed: -6782648765951880356
Simulation 35 seed: 2329996172539762892, L1 agent seed: 4094401061340568415, L2 agent
seed: -7139590729609743879
Simulation 40 seed: -8044238033272326678, L1 agent seed: -4889664622285833917, L2
agent seed: 2305132387050210601
Simulation 45 seed: 1765841120268248596, L1 agent seed: 5939830200348774290, L2 agent
seed: 8352498039242176503
Simulation 50 seed: -7483628206190814146, L1 agent seed: -5901199807367777735, L2
agent seed: -4953801519893314556
Simulation 55 seed: 3964792631180930509, L1 agent seed: -7256341046073327763, L2
agent seed: -1511187976783272592
Simulation 60 seed: 7989216606838704500, L1 agent seed: -6400191989705269235, L2
agent seed: 906626015537785485
Simulation 65 seed: -6706535888773489311, L1 agent seed: 8625520845755199464, L2
agent seed: 8183896729091942626
```

```
Simulation 70 seed: -7683520061444500746, L1 agent seed: 7211111306815403741, L2
agent seed: 2156193192320636203
Simulation 75 seed: 4212970540868612061, L1 agent seed: 2626412600087761926, L2 agent
seed: -4713932784533238189
Simulation 80 seed: 5888678444379199822, L1 agent seed: 506161819105178227, L2 agent
seed: -6263537149442626659
Simulation 85 seed: -216692310953488293, L1 agent seed: 3049492048627163937, L2 agent
seed: 32176859252366458
Simulation 90 seed: 4973636884794263201, L1 agent seed: -8078273577209963985, L2
agent seed: 900946232808125630
Simulation 95 seed: 191641090013900976, L1 agent seed: -5046646292927341189, L2 agent
seed: 4139053336726710829
Simulation 100 seed: -1683528699259121665, L1 agent seed: -6331990536973123276, L2
agent seed: 5173913805948002456

Constant parameters set as follows:
NUM_VALUES: 10
MEANINGS_PER_GENERATION: 200
EROSION_PROBABILITY: 0.001
NUM_TO_ANALYZE: 50
INTELLIGIBILITY_DELAY: 25
INTELLIGIBILITY_THRESHOLD: 0.9
NUM_LANGUAGES: 10
PERCENT_CHANGE: 5

9617025ms to execute
```

As indicated by the final portion of the execution log, the parameter input file, params/values.in, had the following contents:

```
INT    NUM_VALUES                10
INT    MEANINGS_PER_GENERATION   200
DOUBLE EROSION_PROBABILITY       0.001
INT    NUM_TO_ANALYZE            50
INT    INTELLIGIBILITY_DELAY     25
DOUBLE INTELLIGIBILITY_THRESHOLD 0.9
INT    NUM_LANGUAGES             10
INT    PERCENT_CHANGE            5
```

Appendix B

Implementation of Grammar in the Iterated Learning Model

The ILM’s conversion from meaning to signal space is performed via a definite-clause grammar (DCG). Such a grammar consists of one or more rewrite rules in which nonterminal elements are replaced by strings of other nonterminals or terminals, or both. In this context, nonterminals are elements that can be replaced and terminals are elements that cannot be replaced. Thus, elements are terminal if and only if they are part of the signal space. Rules have the form

$$C{:}\ \mu \rightarrow \lambda$$

where C is a category label, μ is a meaning structure, and λ is a sequence of terminals and/or nonterminals. The category label of the rule is based upon the nature of the components contained in μ: if μ only contains an *a*-component, C is A; if μ only contains a *b*-component, C is B, and if μ contains both an *a*- and *b*-component, C is S.

Examples of rules’ application are provided in the form of two simple grammars.

Table B1

Sample Grammars

Grammar 1	Grammar 2
$S{:}\ (a_0, b_0) \rightarrow abc$	$S{:}\ (X,Y) \rightarrow B{:}YA{:}X$
	$A{:}a_0 \rightarrow c$
	$B{:}\ b_0 \rightarrow ab$

(a_0, b_0) yields *abc* (a_0, b_0) yields *abc*

Although Grammar 1 has one rule compared to Grammar 2’s three, they would both produce the string *abc* if prompted to provide a signal for the meaning (a_0, b_0). It is important to note that the meaning structures in the second and third rules of Grammar 2 are not complete meanings. This enables the ILM to produce compositional languages, since they can use rules such as the first rule of Grammar 2 to break apart complex meanings into their constituents.

In the discussion of the induction and invention algorithms, rules such as

$$S: (X,Y) \rightarrow A:YB:X$$

and

$$S: (a_0,Y) \rightarrow cfB:Yp$$

are referred to as general rules, since they contain variables, meaning that they can be used to produce signals for multiple meanings. Similarly, rules involving only specific meanings, such as

$$S: (a_0, b_0) \rightarrow abc$$

are referred to as specific rules.

The algorithms employed for this ILM are designed based on those described by Kirby (2002), which are utilized in Kirby (2001).

Induction Algorithm

The induction algorithm can simplify pairs of rules in three different ways, which are termed double chunking, single chunking, and back-formation. Its structure is detailed below.

Figure C2

Induction Algorithm

IND.1	Select an unordered pair of rules r_1 and r_2.
IND.2	Try to perform double chunking on r_1 and r_2. If this is successful, go to IND.6.
IND.3	Try to perform single chunking on r_1 and r_2. If this is successful, go to IND.6.
IND.4	Try to perform back-formation on r_1 and r_2. If this is successful, go to IND.6.
IND.5	If no induction was performed, stop.
IND.6	For all pairs of rules with the same μ, delete whichever rule has a longer λ
IND.7	Go to IND.1.

Double Chunking

The process for double chunking with two rules r_1 and r_2 is as follows.

Figure C3

Double Chunking

DC.1	If the meanings of r_1 and r_2 differ in more than one position, stop. Otherwise, identify the two components of meaning that differ, calling these m_1 and m_2.	Initial grammar: $S{:}(a_0, b_0) \rightarrow qegjk$ $S{:}(a_0, b_2) \rightarrow qaddjk$
		$m_1 = b_0$ $m_2 = b_2$
DC.2	If the strings of r_1 and r_2 cannot be made the same by removing a non-empty substring from both, stop. Otherwise, identify the shortest such substrings, calling them λ_1 and λ_2.	$\lambda_1 = eg$ $\lambda_2 = add$
DC.3	Create two new rules $C{:}\ m_1 \rightarrow \lambda_1$ and $C{:}\ m_2 \rightarrow \lambda_2$ where C is the appropriate category label given the components that m_1 and m_2 contain.	$B{:}b_0 \rightarrow eg$ $B{:}b_2 \rightarrow add$
DC.4	Replace r_1 and r_2 with a new rule. This rule is identical to both of the former rules, except m_1 and m_2 are replaced with	$S{:}(a_0,\ Y) \rightarrow qB{:}Yjk$

	X and λ_1 and λ_2 are replaced with $C{:}X$. C is a category label as discussed above and X is a variable.	Final grammar: $S{:}(a_0, Y) \rightarrow qB{:}Yjk$ $B{:}b_0 \rightarrow eg$ $B{:}b_2 \rightarrow add$

Single Chunking

The process for single chunking with two rules r_1 and r_2 is as follows.

Figure C4

Single Chunking

SC.1	If the meanings of r_1 and r_2 differ in more than one position or r_1 and r_2 are both specific rules, stop. Otherwise, identify the two components of meaning that differ, calling these m_s and m_g, depending on which rule is specific and which rule is general.	Initial grammar: $S{:}(a_1, Y) \rightarrow B{:}Ytor$ $S{:}(a_1, b_3) \rightarrow ymutor$ $m_g = Y$ $m_s = b_3$
SC.2	If the strings of r_1 and r_2 cannot be made the same by removing a non-empty substring from both of them, stop. Otherwise, identify the shortest such substrings, calling them λ_s and λ_g.	$\lambda_g = B{:}Y$ $\lambda_s = ymu$
SC.3	Create a new rule $C{:}\ m_s \rightarrow \lambda_s$ where C is the appropriate category label given the components that m_1 contains.	$B{:}b_3 \rightarrow ymu$
SC.4	Remove the specific rule.	$S{:}(a_1, b_3) \rightarrow ymutor$ is removed Final grammar: $S{:}(a_1, Y) \rightarrow B{:}Ytor$ $B{:}b_3 \rightarrow ymu$

Back-Formation

The process for back-formation with two rules r_1 and r_2 is as follows.

Figure C5

Back-Formation

BF.1	If the meaning of r_1 is not contained within the meaning of r_2, or vice versa, stop. Otherwise, call the contained meaning *m*. If the meanings are equal, they do not contain each other.	Initial grammar: $S{:}(a_2, b_3) \rightarrow tyvhl$ $A{:}\ a_2 \rightarrow vhl$ $m = a_2$
BF.2	If the string of r_1 is not contained within the string of r_2, or vice versa, stop. Otherwise, call the contained string λ. If the strings are equal, they do not contain each other.	$\lambda = vhl$
BF.3	Replace the rule containing the other rule with a new rule. This rule is identical to the containing rule, except *m* is replaced with *X* and λ is replaced with *C:X*. *C* is the appropriate category label given the components that *m* contains and *X* is a variable.	$S{:}(X, b_3) \rightarrow tyA{:}X$ Final grammar: $S{:}(X, b_3) \rightarrow tyA{:}X$ $A{:}\ a_2 \rightarrow vhl$

Invention Algorithm

The invention algorithm used in the ILM is described below, with μ as a meaning for which the DCG cannot produce a signal.

Figure C6

Invention Algorithm

INV1.	If there are no general rules that differ from μ in exactly one position, generate a random string and go to INV5. Otherwise, find any such general rule and call it *r*. Call the different meaning component in μ m_μ.	Initial grammar: $S{:}(a_0, Y) \rightarrow sgB{:}Ykl$ $B{:}b_0 \rightarrow yot$

		Must generate signal for (a_0, b_1)
		$m_\mu = b_1$
INV2.	Create a temporary rule $C: f \rightarrow s$ where *C* is the appropriate category label given the components that m_μ contains, *f* is a difference flag and *s* is a random string.	$B: f \rightarrow bh$
INV3.	Generate a signal for *μ'*, which is identical to *μ*, except m_μ is replaced by *f*.	$S:(a_0,f)$ generates *sgbhkl*
INV4.	Remove the temporary rule.	$B: f \rightarrow bh$ removed
INV5.	Call the induction algorithm with *μ* and the generated signal.	Induce (a_0, b_1), *sgbhkl*
		Final grammar (via single chunking): $S:(a_0, Y) \rightarrow sgB:Ykl$ $B:b_0 \rightarrow yot$ $B:b_1 \rightarrow bh$