1 Motivation

Second-language acquisition is a much-studied emerging field in linguistics and is of great practical importance in an increasingly global world. However, it is well-recognized that there is a significant variability in the difficulty of learning a new language given a prospective speaker’s native tongue. Some languages are regarded as being very difficult for a native English speaker to learn (Arabic, Korean, etc.) while some are regarded as being easy (French, Spanish, etc.). The three languages ranked as most difficult to learn for native English speakers by State Department study were Japanese, Korean, and Arabic, all of which have orthographic systems significantly distinct from the Latin alphabet used in English. Indeed, having to learn a different orthographic system is often regarded as a significant hurdle on the path to fluency in a second language. As such, it would seem reasonable to examine the properties of these orthographic systems as a possible source of complexity in learning new languages.

The Latin alphabet used by English is particularly well-studied due to its global penetration and acceptance. While one might initially think that Latin script is read from left to right, letter by letter, scanning the entire string of characters, research in the area suggests otherwise. In fact, legibility of text is dependent on a complex set of factors including the shape of a word, spacing between characters, and spacing between lines, among others. Perhaps most interesting is that one does not in fact need the entirety of a written word to be able to recognize it. For Latin script, one may read and comprehend text with only the top half of the words. Indeed, we may make a basic demonstration of this phenomenon:

Figure 1: Guess the string!

---


But why does this phenomenon exist in the Latin orthographic system? It would seem as though more information is contained in the top portion of letters as they are canonically rendered. That is to say, it seems as though the Shannon entropy of Latin characters is positionally dependant - higher points in the letter have higher entropy and therefore characters carry more information at those loci. As we move forward in our proposal, we will refer to characters as graphemes. A grapheme is the smallest component of an orthographic system which can be combined with others to make words. We will need the notion of a grapheme as we devise an analytic framework sufficiently general to encapsulate all known written languages.

We hypothesize that different orthographic systems will have different fixation points based on how the entropy of text varies with visual position. We further hypothesize that knowledge of the fixation line of an orthographic system can help language learners speed second-language acquisition by training them to look along the path of highest information content.

2 Experimental Design

We propose an experimental design split into two distinct stages: evaluation of information content in a set of orthographic systems, and path-directed reading for intermediate learners of a language. In all experiments, we plan to examine the orthographic systems of: English, Hindi, Korean, Arabic, and Chinese, adding others if time allows.

2.1 Information Content Profiling

We hope to follow a standardized procedure for information content profiling:

1. Determine general bounding size for all graphemes.
2. Render all graphemes in the target orthography with standardized font, size, and centering.
3. Determine weighting for each grapheme by their frequency in a corpus text.
4. Overlay normalized grapheme renderings weighted by frequency in a corpus text and build a map of most highly variable areas.
2.2 Path-directed Reading

Here we hope to prove that the data obtained during profiling is of practical use and relevance. We will choose study participants who are currently in "L2" of a targeted language at Yale University. We hope that path-directed reading will allow language learners to learn a new language faster.

1. Generate test strings of language to be read by students.
2. Automatically annotate prospective text at location of highest information.
3. Present battery of annotated and non-annotated tests to student and time his or her response.

Figure 3: Differential paths in Hangul (Korean) and Latin script

3 Deliverables

1. Code to render all possible graphemes of a general language
2. Code to read corpus, weight renderings and generate entropy maps
3. Code using entropy maps and prospective text to generate path-direction annotations
4. Code to accurately determine time from presentation of stimulus to processing.
5. Statistical analysis of path-directed reading results

4 Conclusion

Ultimately, I hope that this research will be useful to those learning new languages and struggling with new orthographic systems. I believe that this research will also inform an important and widely studied topic from a novel perspective, and provide a computable measure with which to characterize orthographic systems.