Creating an artificial language for mitigating the Internet Information Retrieval recall problem

XML

1. Introduction

Artificial languages (AL) come in many flavours and have been created for many purposes, consider for instance (Sanders 2016: 2) and are known at least since the 11th century (Higley 2007: 4-7). We create one in order to mitigate the recall problem of Internet Information Retrieval for Low Resource Languages (LRL). This problem refers to the fact that for any LRL, which one seeks to build a corpus for, compare (Scannell 2007: 4), one does not know how much of all available content on the web in the LRL is covered by what one has found. We created a corpus in an AL calling it Jucearm which we then posted in various places on the web. The AL we created is a so-called *a posteriori language* meaning that we took an existing language and altered it before applying a non-trivial and non-traceable letter substitution schema. We then place the contents from our corpus in bits in various social networks and on homepages. 5 students have then been provided with a noisy wordlist of Jucearm imitating a realistic situation for many LRL retrieval attempts and a short grammatical sketch and were told to search all available content on the web. We release the corpus alongside this publication. [1]

2. Creating Jucearm

We chose an a posteriori approach because the statistics of natural language are already inherent in the basis. And another advantage is that there are many resources and corpora for a widely used language on which to build the language. Our approach was to look for a major language that exhibits largely average statistical properties at various linguistic levels. Therefore, we considered WALS [2] for large web languages [3] in the categories: Consonant Inventories (Maddieson 2013a), VowelQuality Inventories (Maddieson 2013c) and Syllable structure (Maddieson 2013b). [4]

Table 1. General aspects of language structure for applicable large web languages. Table 2. Queries and results for Jucearm on the “virgin web” (before Jucearm).

Language	Vowel Inventory	Consonant Inventory	Syllable Structure
English	Large (7-14)	Average	Complex
Russian	Average (5-6)	Moderately large	Complex
Spanish	Average (5-6)	Average	Moderately complex
Turkish	Large (7-14)	Average	Moderately complex
Persian	Average (5-6)	Average	Complex
French	Large (7-14)	Average	Complex
German	Large (7-14)	Average	Complex
Japanese	Average (5-6)	Moderately small	Moderately complex
Portuguese	Large (7-14)	Average	Moderately complex
Vietnamese	Large (7-14)	Average	Moderately complex

We chose Spanish. We downloaded and merged all Spanish corpora from Universal Dependencies 2.6. [5] Before substituting letters, we applied the following changes:

1. We substituted the upside-down question mark characteristic for Spanish orthography by a question particle, which we called ’hal’ as its Arabic equivalent. However, this was applied to content and yes/no questions.
2. We substituted all feminine definite articles in either singular or plural by their masculine counterpart. Although this transformation did not change the complete gender system, it is statistically significant at least in theory.
3. We substituted the synthetic future with the -ra type endings by an analytic one with the verb volver. Thus ’tu comerás’ becomes ’tu vuelves comer’.
4. Finally, we added a peculiar feature to the source language, which is not very widespread in European languages, at least not for the expression of grammatical properties: partial reduplication. For superlatives of Adjectives or Adverbs, which were continuous tokens as ’el más X’, we substituted by the initial sequence until the (first or if present) second vowel, which resulted in a full reduplication if the token was short, such as *bajobajo*. Examples: *descodesconocido, impoimportante, evievidente s*.

After this, we applied a letter mapping with a sum total of 36 transformations, including three 2:1 and three 1:2 transformations. Spanish isthus not trivially reconstructible. If one would expect text in the AL to be encrypted, then a letter frequency analysis independent of the letters could reconstruct it otherwise.

Upon development, we created more than one mapping and each time consulted the statistics of the generated text until we found the result to be normal with respect to statistics. The tools we used to analyze the generated texts were:

1. Analysis of the distribution of letters with R in comparison to a variety of languages.
2. Zipfplots created with R and compared to Zipfplots of other languages (Montemurro 2011).
3. We combined usually three words (low frequency, mid-frequency and high-frequency words) and noted the primary languages of the first 20 pages found in a web query. The desired result should i) be diverse and ii) not feature a lot of Spanish.

Table 2. Queries and results for Jucearm on the “virgin web” (before Jucearm).

Queries	Results
ba (HF) ajs (HF) az (HF)	English 12, German 6, Polish 1, Codes 1
az (HF) sjemdu (MF) wepez (LF)	no results
zovasj (MF) ba (HF) asj (HF)	English 5, German 3, Codes 2, Turkish 1
az (HF) sjemdu (MF) zovasj (MF)	no results
sjebu (MF) wepez (LF) azvamnuz (LF)	no results

In the grapheme system of the AL appear grapheme sequences which seem difficult to pronounce and unusual, such as <cj> or <sj>. However, these grapheme sequences constitute graphemic units as parts of writing systems and could for instance represent palatalized sounds in our AL. For example, <cj> is a part of the Friulian writing system and represents the voiceless palatal stop /c/ (Miotti 2002) or <sj> which consitutes the voiceless velopalatal fricative /ɧ/ in Swedish (Holmes / Hinchliffe 2013). This means that the language is readable applying some idiosyncratic graphemic unitto phoneme rules such as <sj> /ʃ/ as many languages do possess. An example of the original Spanish and our correspondent a posteriori AL:

Spanish: Es como si los hubieran estado golpeando contra el pared durante horas.

Jucearm: Uz sunu zo sjuz figoamez azpebu dusjwaezbu suzpme asj wemab bimezpa fumez.

3. Named Entities

Transforming a text in the above-described way would also affect Named Entities(NE), which can be orthographically much more complex than average words of a language and often include abbreviations/ acronymes. A text without Nes would be unusual and a text with almost only language internal NEs and no loaned ones maybe even more so.

We extract a list of NEs using spaCy’s Named Entity Recognizer (NER). [6] The data-set we used for the extraction consists of Wikipedia dumps of 14 different languages: Afrikaans, Albanian, Basque, Danish, Estonian, French, Italian,Croatian, Lithuanian, Polish, Portuguese, Romanian, Swedish and Turkish. We then used the program wiki-extractor [7] to extract only text data from the Wikidump. Using spaCy and individually trained models, we extracted overall 8.000 tokens of Named Entities.

We then substituted all the NE tagged instances from the UD base corpus by one random entry of the NER list. Sometimes we slightly adapted the names to typical Jucearm letter sequences. We classified the result,the final corpus, with Facebooks fastText [8] which gave Slovenian as a result.

4. Migration into the internet

The (main) goal of this paper is to imitate/emulate a situation of a typical LRL. We use our experience in the work with LRLs on the web (Hoenen et al. 2020) to find suitable places. Yet, we want to be legally certain to put our AL in specific places on the internet, where it is first legal/allowed and second does no damage. In order to verify the validity of both mentioned aspects, we looked at the general terms and conditions as well at all privacy policies on every page in which we have placed our text. Furthermore we made sure, that it is possible to delete each placed text after the experiment. We placed some content on our homepages (2 domains, with a pdf also placed), placed some content on social media and bid a ‘book in Artificial Language’ with an excerpt of Jucearm on a vendor platform. We also placed a genuine sentence in the AL on Wikipedia in a fitting article where it persisted in Ukrainian.

5. Live-Experiment

We created a “starter kit” in the An Crúbadán format [9] with a word list and a short grammatical sketch. Based on the starter package, we instructed 5 participants to search for the placed contents in the AL. The results of the experiment showed that all participants had found at least one of the subpages on our own websites; four had also found more than one piece of content. Four participants found the text in the Wikipedia article and three found our classified ad in the AL. In the end, none of the participants found the entries on Social Media. All participants used search engines (primarily Google) for the experiment, by using queries of multiple terms, single words, or the name of the AL itself.

6. Discussion and Conclusion

Since most probably for no living medium-size LRL on the web all places of content are known with certainty, approaching the quality of the simulation must remain controversal. Despite being only a small start, the results are however encouraging and could point to modern corpus crawling for LRLs being quite effective concerning recall.

Appendix A

Bibliography

1. **Ackerlind, Sheila R.** / **Jones-Kellogg, Rebecca** (2011): *Portuguese: A reference manual.* University of Texas Press.
2. **Higley, Sarah** (2007): *Hildegard of Bingen’s unknown language: An edition, translation, and discussion*. Springer.
3. **Hoenen, Armin** / **Koc, Cemre** / **Rahn, Marc D.** (2020): “A manual for webcorpus crawling of low resource languages.” In: *Umanistica Digitale*, 4(8).
4. **Holmes, Philip** / **Hinchliffe, Ian** (2011): *Swedish: A comprehensive grammar.*
5. **Maddieson, Ian** (2013a): “Consonant inventories.” In: Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.
6. **Maddieson, Ian** (2013b): “Syllable structure.” In: Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.
7. **Maddieson, Ian** (2013c): “Vowel quality inventories.” In: Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.
8. **Mateus, Maria H.** / **d’ Andrade, Ernesto** (1998): “The syllable structure in european portuguese.” In: *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada*, 14(1):13–32.
9. **Miotti, Renzo** (2002): “Friulian”. In: *Journal of the International Phonetic Association* 32(2):237–247, 2002.
10. **Montemurro, Marcelo A.** (2001): “Beyond the zipf–mandelbrot law in quantitative linguistics.” In: *Physica A: Statistical Mechanics and its Applications*, 300(3-4):567–578,.
11. **Sanders, Nathan** (2016): “Constructed languages in the classroom.” In: *Language* 92(3):e192–e204.
12. **Scannell, Kevin P.**(2007): “The Crúbadán Project: Corpus building for under-resourced languages.” In: *Building and Exploring Web Corpora: Proceedingsof the 3rd Web as Corpus Workshop*, volume 4:5–15.

Notes

1. https://github.com/ArminHoenen/JuCeArm
2. https://wals.info/
3. https://de.statista.com/statistik/daten/studie/2961/umfrage/anteil-der-verbreitetsten-sprachen-im-internet/
4. If no entry in WALS, we considered Ackerlind / Jones-Kellogg (2011) and Mateus / Andrade (1998).
5. universaldependencies.org
6. https://spacy.io
7. https://github.com/attardi/wikiextractor
8. https://fasttext.cc/
9. http://crubadan.org/

Armin Hoenen (hoenen@em.uni-frankfurt.de), Goethe Universität Frankfurt, Germany and Cemre Koc (cem_koc@icloud.com), Goethe Universität Frankfurt, Germany and Julian Hasche (julian.hasche@gmail.com), Goethe Universität Frankfurt, Germany