



Cthulhu Hails from Wales

N-gram Frequency Analysis of R'lyehian

Vít Novotný  and Marie Stará 

Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
{413827,409729}@mail.muni.cz

A'bstract R'lyehian is a unique fictional language penned by the prolific 20th century horror fiction author H. P. Lovecraft. Prior work in the area of the Lovecraftian mythos has not yet studied the similarities between R'lyehian and natural languages, which are crucial for determining its true origins. We produced a comprehensive wordlist of R'lyehian and used open-source *N*-gram-based language identification tools to find the most similar natural languages to R'lyehian. From the comprehensive wordlist, we also constructed a frequency table of all unigraphs and digraphs in R'lyehian. We show that R'lyehian is most similar to Celtic languages, which lays grounds for our hypothesis that R'lyeh, where Cthulhu lies dreaming, might be a place in Wales. Our frequency tables will prove a useful resource for future work in the area of the Lovecraftian mythos.

K'eywords: H. P. Lovecraft, language identification, *N*-grams, R'lyehian

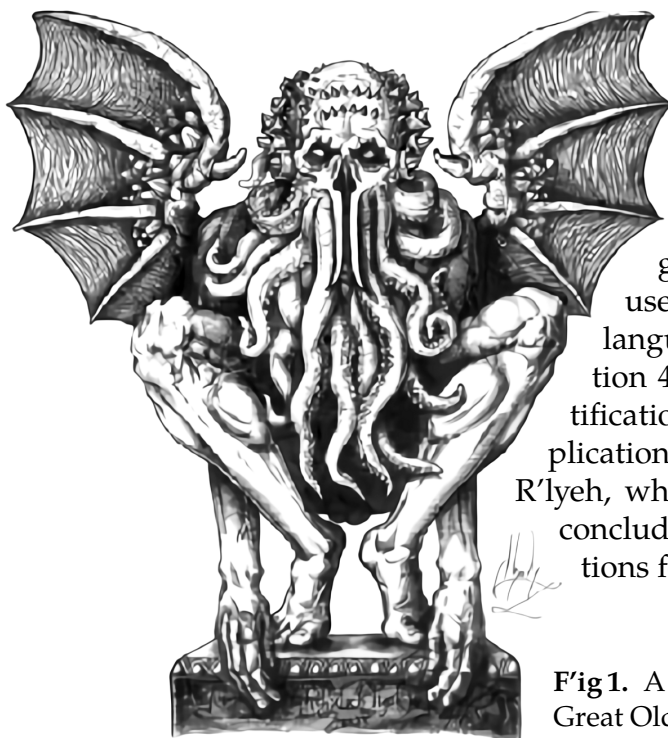
1 I'ntroduction

H. P. Lovecraft is regarded as one of the most influential authors of the 20th-century horror genre. R'lyehian is a fictional language spoken by ancient cosmic dieties (the *Great Old Ones*, see F'igure 1) in Lovecraft's 1926 short story *Call of Cthulhu* [7] and in his later work. Below is an example sentence in R'lyehian:

Ph'nglui mglw'nafh Cthulhu R'lyeh wgah'nagl fhtagn.
In his house at R'lyeh dead Cthulhu lies dreaming.

Prior work in the area of the Lovecraftian mythos has neglected the similarities of R'lyehian to natural languages and has focused mainly on Lovecraft's use of English. [5,13] Since R'lyehian has been romanized, it lends itself to character *N*-gram frequency analysis and therefore also language identification.

Prior work has not determined the exact location of the sunken city of R'lyeh. Lovecraft (1928) [7] places R'lyeh at 47°9'S 126°43'W in the southern Pacific Ocean, whereas Derleth (1952) [4], a correspondent of Lovecraft, places R'lyeh at 49°51'S 128°34'W. By identifying the most similar natural languages to R'lyehian, we hope to discover the true location of the resting place of the Old One Cthulhu.



Our work is structured as follows: In Section 2, we briefly discuss R'lyehian and its phonology. In Section 3, we describe our romanized wordlist and the open-source N -gram-based language identification tools that we used to find the most similar natural languages to written R'lyehian. In Section 4, we present the language identification results and discuss their implications for locating the sunken city of R'lyeh, where Cthulhu lies dreaming. We conclude in Section 5 and suggest directions for our future work.

F'ig 1. A monolith of Cthulhu, one of the Great Old Ones (Alexander Liptak, cc by). [3]

2 R'lyehian

R'lyehian, also known as Cthuvian, is a language created by H. P. Lovecraft for his 1926 short story *Call of Cthulhu* [7]. Unlike some other fictional languages, such as the J. R. R. Tolkien's Elvish languages,¹ or Marc Okrand's Klingon [10] from the Star Trek universum, Lovecraft's R'lyehian appears only in fragments and has no comprehensive vocabulary or grammar.

Below, we list a few facts known about R'lyehian:

- It is (supposed to be) unpronounceable for humans. [12]
 - As it uses a number of different prefixes and suffixes, it can be classified as a synthetic language.² Unfortunately, not enough data exist to subclassify it more accurately as either an agglutinative or a fusional (inflected) language.
 - It makes no distinction between the past and the future, only the present and the non-present, [9] since the Old Ones exist at all times simultaneously.
 - It does not distinguish parts of speech and has free word order. [11,1]
 - It is written in hieroglyphics, see e.g. the text on the pedestal in F'igure 1. [9]
- The romanized spelling reflects how English speakers captured the speech.³

¹ Tolkien's Sindarin was partially based on Welsh, which Tolkien discusses in his 1931 essay *A Secret Vice* [14].

² https://en.wikipedia.org/wiki/Synthetic_language

³ “[The word Cthulhu is] supposed to represent a fumbling human attempt to catch the phonetics of an *absolutely non-human* word.” (Lovecraft, 1976) [8, pp. 10–11]

Some useful insights about R'lyehian can be found in the work of Robinson (2010) [12] describing the names (teratonyms⁴) used by H. P. Lovecraft. Robinson describes the features Lovecraft used to make the language seem unpleasant and harsh as well as the influence of other languages (Arabic, Hebrew, and fragments of African languages) on teratonyms. Some of their conclusions can be applied to the language of R'lyehian as a whole.

The intentional strangeness of R'lyehian language was, according to Robinson, produced at three levels:

1. individual sounds,
2. sound combinations, and
3. word-forms.

At the first level, the strangeness was produced by clustering consonants atypical for English, such as the aspirated consonants or various nasal combinations, e.g. *bn*, *mn*, *mt*, *mth*, or *pn*.

At the second level, the unpronouncibility was produced similarly to the first level by creating clusters unnatural for English or by using clusters that appear in English, but placing them “in patterns or positions that run contrary to its phonotactics”. For example: beginning a syllable with a cluster that usually appears at the end of English words, such as *pth* in *depth*.

As for the third level, it can be stated simply by looking at the words in R'lyehian that it seems and sounds unnatural and strange. To achieve this goal, Lovecraft used low (*a*) and back (*u*, *o*) vowels and consonants that are perceived as harsh and dissonant.

2.1 P'ronunciation

There are no clear rules for pronouncing R'lyehian. To the best of our knowledge, Lovecraft himself described merely the pronunciation of the name *Cthulhu*:

“The actual sound — as nearly as human organs could imitate it or human letters record it — may be taken as something like *Khûl'hloo*, with the first syllable pronounced gutturally and very thickly. The *u* is about like that in *full*; and the first syllable is not unlike *klul* in sound, since the *h* represents the guttural thickness. The second syllable is not very well rendered — the *l* being unrepresented.” (Lovecraft, 1976) [8, p. 11]

3 M'ethods

To identify the most similar natural languages, we required a corpus or a wordlist of R'lyehian and an *N*-gram-based language identification tool with pre-trained models for natural languages. In this section, we present our comprehensive wordlist and our frequency table of all unigraphs and digraphs in R'lyehian, and the language identification tools that we used in our experiment.

⁴ Names of monsters: *terato* (monster) + *nym* (name)

3.1 R'lyehian wordlist

Due to the sparse occurrences of R'lyehian in Lovecraft's work, we decided against producing a R'lyehian corpus. Instead, we collated two online resources [11,1] into a comprehensive wordlist that we show below in alphabetical order:

1. <i>ah</i>	25. <i>grah'n</i>	49. <i>n'gha</i>	73. <i>tharanak</i>
2. <i>athg</i>	26. <i>h'ehye</i>	50. <i>n'ghft</i>	74. <i>thflthkh'ngaha</i>
3. <i>bug</i>	27. <i>hafh'drn</i>	51. <i>naf'lthagn</i>	75. <i>throd</i>
4. <i>bugg-shoggog</i>	28. <i>hai</i>	52. <i>nglui</i>	76. <i>uaaah</i>
5. <i>cf'ayak</i>	29. <i>hastur</i>	53. <i>nilgh'ri</i>	77. <i>uh'e</i>
6. <i>cf'tagn</i>	30. <i>hlirgh</i>	54. <i>nog</i>	78. <i>uln</i>
7. <i>chtenff</i>	31. <i>hrii</i>	55. <i>nw</i>	79. <i>ulnagr</i>
8. <i>cthugha</i>	32. <i>hupadgh</i>	56. <i>ooboshu</i>	80. <i>vugtlag'n</i>
9. <i>cthulhu</i>	33. <i>iä</i>	57. <i>orr'e</i>	81. <i>vugtlagln</i>
10. <i>ebumna</i>	34. <i>ilyaa</i>	58. <i>ph'nglui</i>	82. <i>vulgtlagln</i>
11. <i>ee</i>	35. <i>k'yarnak</i>	59. <i>ph'nglui</i>	83. <i>vulgtm</i>
12. <i>ehye</i>	36. <i>kadishtu</i>	60. <i>phlegeth</i>	84. <i>vulgtmm</i>
13. <i>ep</i>	37. <i>kn'a</i>	61. <i>r'luh</i>	85. <i>wgah'n</i>
14. <i>farnomi</i>	38. <i>li'hee</i>	62. <i>r'lyeh</i>	86. <i>wgah'nagl</i>
15. <i>fhagn</i>	39. <i>llll</i>	63. <i>ron</i>	87. <i>y'bthnk</i>
16. <i>fhthagn-ngah</i>	40. <i>lloig</i>	64. <i>s'uhn</i>	88. <i>y'hah</i>
17. <i>fm'latgh</i>	41. <i>lw'nafh</i>	65. <i>sgn'wahl</i>	89. <i>ya</i>
18. <i>fomalhaut</i>	42. <i>mg</i>	66. <i>shagg</i>	90. <i>ygnailh</i>
19. <i>ftaghu</i>	43. <i>mglw'nafh</i>	67. <i>shogg</i>	91. <i>yog-sothoth</i>
20. <i>geb</i>	44. <i>mnahn'</i>	68. <i>shtunggli</i>	92. <i>yuggoth</i>
21. <i>gnaiih</i>	45. <i>n'gai</i>	69. <i>shugg</i>	93. <i>zhro</i>
22. <i>gof'nn</i>	46. <i>n'gha'ghaa</i>	70. <i>sll'ha</i>	
23. <i>goka</i>	47. <i>n'gha-ghaa</i>	71. <i>stell'bsna</i>	
24. <i>gotha</i>	48. <i>n'grkdl'lh</i>	72. <i>syha'h</i>	

From the wordlist, we extracted the affixes of R'lyehian:

1. <i>-agl</i>	5. <i>-og</i>	9. <i>c-</i>	13. <i>ng-</i>
2. <i>-agn</i>	6. <i>-or</i>	10. <i>h'-</i>	14. <i>nnn-</i>
3. <i>-agr</i>	7. <i>-oth</i>	11. <i>na-</i>	15. <i>ph'-</i>
4. <i>-nyth</i>	8. <i>-yar</i>	12. <i>nafl-</i>	16. <i>y-</i>

From the wordlist, we also constructed a frequency table of all unigraphs and digraphs in R'lyehian in Table 1. Our table shows that R'lyehian consists of 7 vowels and 28 consonants, including 11 digraphs mostly created by the consonant *+h*, which changes the pronunciation of the first consonant.

3.2 Language identification

In this section, we describe the open-source language identification tools that we used in our experiment. Our selection is based on the survey of Jauhiainen et al. (2019) [6]. We report the top three languages identified by the tools.

T’able 1. Frequencies of all unigraphs and digraphs in R’lyehian extracted from our comprehensive wordlist. We categorize the unigraphs into consonants and vowels.

Unigraphs		Digraphs	
Consonants	Vowels		
<i>g</i>	9.06%	<i>a</i>	12.33%
<i>n</i>	7.90%	<i>’</i>	7.71%
<i>l</i>	7.51%	<i>u</i>	5.59%
<i>h</i>	5.39%	<i>o</i>	4.05%
<i>r</i>	3.47%	<i>i</i>	3.85%
<i>t</i>	3.08%	<i>e</i>	3.47%
<i>f</i>	2.31%	<i>ä</i>	0.19%
<i>y</i>	2.31%		
<i>m</i>	1.93%		
<i>k</i>	1.73%		
<i>s</i>	1.54%		
<i>b</i>	1.35%		
<i>w</i>	1.16%		
<i>d</i>	0.96%		
<i>v</i>	0.96%		
<i>c</i>	0.77%		
<i>p</i>	0.39%		

T’able 2. The top three closest natural languages to R’lyehian identified by three different language identification tools. Celtic languages are *emphasized*.

Tools	Languages
TextCat	<i>Scots, Manx, Welsh</i>
CLD2	<i>Irish, Croatian, Sesotho</i>
LangDetect	<i>Somali, Indonesian, Welsh</i>

TextCat In their seminal work, Cavnar et al. (1994) [2] described the *out-of-place* *N*-gram-based language identification method, which is implemented by the open-source TextCat tool.⁵ TextCat contains models for 69 natural languages.

CLD2 Compact Language Detector 2⁶ (CLD2) is the language identifier from the Google Chrome web browser. For Unicode blocks that map one-to-one to detected languages, CLD2 uses simple rules. For others, CLD2 uses a Naive Bayes classifier on character *N*-grams. CLD2 contains models for 160 natural languages.

LangDetect LangDetect⁷ is a language identifier that uses a Naive Bayes classifier on character *N*-grams. Like CLD2, LangDetect applies a number of normalization heuristics to the input text. LangDetect supports 55 natural languages.

4 R’esults

T’able 2 places R’lyehian closest to Celtic languages (Scots, Manx, Welsh, and Irish) with Welsh being the most frequent among the top three closest languages. As a result, we hypothesize that R’lyeh might be the Caldey Island in Wales at 51°38’N 4°41’W, where hooded monks observe Celtic rites and make offerings of the darkest of chocolates to the slumbering Cthulhu.

⁵ <https://www.let.rug.nl/~vannoord/TextCat/>

⁶ <https://pypi.org/project/cld2-cffi/>

⁷ <https://pypi.org/project/langdetect/>

5 Conclusion

Although Lovecraft's fictional language of R'lyehian is purposely distinct from natural languages, our results suggest that R'lyehian was inspired, either consciously or subconsciously, by the Celtic language of Welsh.

Future work should compare the phonology of Welsh and R'lyehian using our comprehensive wordlist and frequency table of all unigraphs and digraphs, and expand our wordlist by manning a Wales expedition to interview Cthulhu.

Acknowledgments. First author's work was funded by the South Moravian Centre for International Mobility as a part of the Brno Ph.D. Talent project.

References

1. Admin of Naguide.com: Call of Cthulhu – R'lyehian Language Guide (November 2018), <https://www.naguide.com/call-of-cthulhu-rlyehian-language-guide/>
2. Cavnar, W.B., Trenkle, J.M., et al.: *N-gram-based text categorization*. In: Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval. vol. 161175 (1994)
3. Cole, D.R., et al.: Cthulhuic Literacy: Teaching Secondary English with a Dose of Lovecraft. *English in Australia* **49**(1), 72 (2014)
4. Derleth, A.: The Black Island. *Weird Tales* (January 1952)
5. Jamneck, L.: Tekeli-li! Disturbing Language in Edgar Allan Poe and H. P. Lovecraft. *Lovecraft Annual* (6), 126–151 (2012), <https://www.jstor.org/stable/26868454>
6. Jauhiainen, T.S., Lui, M., Zampieri, M., Baldwin, T., Lindén, K.: Automatic language identification in texts: A survey. *JAIR* **65**, 675–782 (2019)
7. Lovecraft, H.P.: The Call of Cthulhu. *Weird Tales* (February 1928)
8. Lovecraft, H.P.: *Selected Letters [V] 1934–37*. Arkham House (1976)
9. Luethke, K.: *Fathoming the Unknown: A Divulge into H. P. Lovecraft's Use of Linguistic Phonology and Entomology in Relation to Cosmic Horror and the Cthulhu Mythos*. The Luethke Company (2014)
10. Okrand, M.: *The Klingon Dictionary: The Official Guide to Klingon Words and Phrases*. Simon and Schuster (1992)
11. Roadagain, R., Haq, A., et al.: R'lyehian (November 2020), <https://lovecraft.fandom.com/wiki/R'lyehian>
12. Robinson, C.L.: Teratonymy: The Weird and Monstrous Names of HP Lovecraft. *Names* **58**(3), 127–138 (2010), <https://doi.org/10.1179/002777310X12759861710420>
13. Spencer, H.: *Semantic Prosody in Literary Analysis: A Corpus-based Stylistic Study of H. P. Lovecraft's stories*. Master's thesis, University of Huddersfield (2011)
14. Tolkien, J.R.R.: *The Monsters and the Critics, and Other Essays*. George Allen and Unwin (1986)