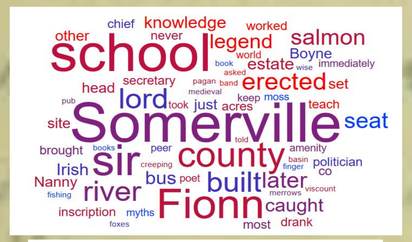
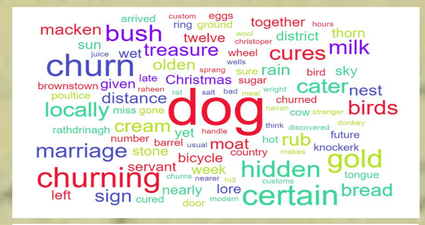
Using statistics to investigate changing use of language in Irish Primary students' writing after 80 years.



Word cloud of frequently used words in 2018 data.



Word cloud of frequently used words 1930's Dúchas data.

Introduction

The language and words we use tells us a lot about our society, our social interactions and what is important in our lives.
As time goes on and population changes so does the components of their language.

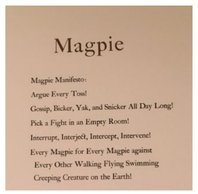
Why did we do this project

The idea for this project came from the folklore stories written by Kate's Great Grandaunt Julia Clancy in 1937. These stories became available on line from Dúchas.

- Dúchas is an Irish online website that contains over 740,000 pages of folklore written by Primary school children from 5,000 different school around Ireland written between 1937-39.
- The Duchas projects objective was to digitize the National Folklore collection so that the public has online access to the materials and a data management system becomes available.

What did we hope to achieve

- We set out to use math's and statistics to look for patterns and word usage in the writings of primary school children in the late 1930's and compare them to the data in similar pieces of writing we have collected from schoolchildren in the same geographical location (Kentstown, Co.Meath) in 2018.
- We decided to use Natural language toolkit and online software to generate a count of word occurrences, to establish the most frequently used words and look for words that give away a writers gender and look at the change over time. We were also interested in any differences in readabilty level between the 1930's and 2018 stories. We used a Flesch-Kinkaid English readability test to compare them.
- Ben Blatt is a statistician who uses maths to look at patterns in books and classic novels. We used similar methods to Blatt to look for patterns in our data. "The written word and the word of numbers should not be kept apart" Blatt,B.,*2018:Nabokovs favourite word is mauve.* London. Simon&Schuster.
- Anne Bayetto from Flinders University investigated high frequency words in young children's writing in Australian students' writings from 2007 and then from 2017. We planned to compare our data to her findings.
- Many nature words like Magpie, dandelion, buttercup and acorn have been taken out of the Oxford Children's dictionary because children don't use them as much anymore. We looked to see if any of these words appear in the primary school students writing now or in the writings from 80 years ago.
- Robert MacFarlane and Jackie Morris published a book using some of these remover or "lost words". The book uses poems and pictures about the missing words like Magpie and Dandelion to show "the spirits of their subjects".(1)



1. Macfarlane , R. & Morris, J., 2017. The Lost words. London: Hamish Hamilton.

Experimental methods:

1. Transcribe the folklores on Dúchas. https://www.duchas.ie/en
2. Visit Kentstown National school and collect Folklores written by present day students.
3. Randomly select a number of folklores from the Kentstown Dúchas collection on the calculator to match the number collected in Kentstown school 2018.
4. Use Python and the Natural language tool kit to find the frequency of the words used.

```
File  Edit  Format  Run  Options  Window  Help
import nltk

data = open("allKbakhdata.txt", "r", encoding="utf8")

tokens = nltk.word_tokenize(data.read().lower())

freq = nltk.FreqDist(tokens)

print(freq.items())
```

5. Analyze data using the Text analyser https://www.online-utility.org/text/analyzer.jsp
6. Use the following site to get Flesch Kincaid score for each folklore https://www.onlineutility.org/english/readability_test_and_improve.jsp
7. Use Excel and the results from the Flesch-Kincaid to do a t-Test comparing the 2018 and the 1930 results.
8. Make Word clouds of the most frequently used words in the 2018 and the 1930 folklores for each gender using the site https://worditout.com/word-cloud/create

9. Calculate Ratios for words that one gender uses frequently compared to the other gender in the 1930's and 2018 data.
10. Compare words used in the 2018 folklores that were not used in the 1930 folklores or words used in the1930's and not used in 2018.
11. Use Python to check for the lost/ removed from the Oxford Children's dictionary in 2007 in all the 1930 Kentstown folklores on Dúchas.
12. Check the Kentstown folklores for those 'lost' words.

Results:

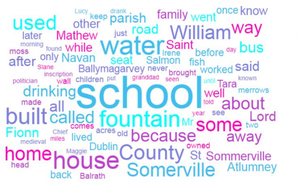


Fig:1 Frequency of words used by girls in 2018

Fig:2 Frequency of words used by girls in the 1930's

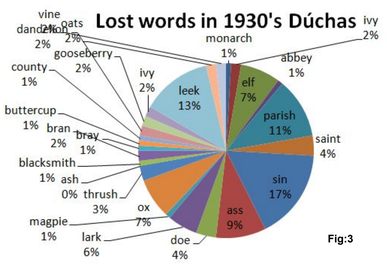


Fig:3

Results:

Schools related words only in 2018 folklores

School	0.4195
Knowledge	0.1291
Teach	0.0968
Book	0.0645
Books	0.0645
Classroom	0.0487
Schools	0.0323
Teacher	0.0161
Students	0.0323
Pupils	0.0323
Classmates	0.0161

Fig: 4

Generic folklore related words only in 2018 folklores

Fionn	0.3227
Salmon	0.1775
Boyne	0.1129
Myths	0.0827
River	0.2259
Fianna	0.0968
Cumhall	0.0484
Burial	0.0161
Tombs	0.0161
Newgrange	0.0323

Fig:5

Job (profession) related words only in 2018 folklores

Secretary	0.3227
Politician	0.1775
Poet	0.1129
Teacher	0.0827
Soldier	0.2259
Vicar	0.0968

Contracted form words only in 2018 folklores

Didn't	0.3227
Don't	0.1775

Fig:6

Formal title related words only in 2018 folklores

Sir	0.3388
Lord	0.2259
Lady	0.1613
Viscount	0.0645
Chief	0.0968

Fig:7

Words about themselves only in 2018 folklores

I	0.2743
Am	0.0484
Me	0.0645

Fig:8

Nature related words only in the Duchas collection

Dog	0.3074
Bush	0.1383
Birds	0.1229
Nest	0.0922
Thorn	0.0768
Birds	0.0615
Cow	0.0615
Rat	0.0461
Donkey	0.0461
Snail	0.0307
Robin	0.0307
Dandelion	0.0154
Magpie	0.0154

Fig:9

Food related words only in the Duchas collection

Butter	0.1537
Milk	0.1229
Bread	0.1076
Cream	0.1076
Sugar	0.0615
Eggs	0.0615
Garlic	0.0461
Juice	0.0461
Meal	0.0461
Vinegar	0.0461
Mustard	0.0307
Pudding	0.0307
Salt	0.0461

Fig:10

Farming related words only in the Duchas collection

Churned	0.0615
Churning	0.169
Churns	0.0461
Churn	0.1844
Wool	0.0461
Shepherd	0.0307
Spinning	0.0307

Fig:11

Weather related words only in the Duchas collection

Weather	0.2613
Rain	0.0922
Sun	0.0768
Wet	0.0768
Sky	0.0768
Cloud	0.0307

Fig:12

Gender word use comparison 2018

Gender	Girl (New)	Boy (New)	Ratio (New) Girl : Boy
House	0.3965	0.0828	4.79 to 1
Home	0.2908	0.0414	7.02 to 1
Family	0.185	0.1656	1.12 to 1
School	0.6609	0.0414	15.9 to 1

Fig:13

Gender word use comparison 2018

Gender	Girl (New)	Boy (New)	Ratio (New) Girl : Boy
Fish	0.1586	0.4969	1 to 3.133
Hills	0.0264	0.6625	1 to 25.095
River	0.0793	0.4555	1 to 5.744
Estate	0.0264	0.2899	1 to 10.981
Viscount	0	0.1656	Boy only
Lady	0.1057	0.2484	1 to 2.35

Fig:14

Gender word use comparison 2018

Fig:15

Gender word use comparison 2018

Fig:16

Gender word use comparison 1930's

Gender	Girl (Old)	Boy (Old)	Ratio (Old) Girl : Boy
Eggs	0.0959	0	Girls only
Cream	0.1199	0.0856	1.4 to 1
Butter	0.1918	0.0856	2,24 to 1
Churn	0.2158	0.1284	1.68 to 1

Fig:17

Gender word use comparison 1930's

Fig:18

Gender word use comparison

Gender	Girl (New)	Boy (New)	Ratio (New) Girl : Boy	Girl (Old)	Boy (Old)	Ratio (Old) Girl : Boy
She	0.819	0.4555	1.8 to 1	0.2637	0.1284	2.05 to 1
He	1.2424	1.2422	1 to 1	0.7193	1.5411	1 to 2.14
Mum	0	0	0	0	0	0
Her	0.4494	0.3313	1.36 to 1	0.1918	0.1712	1.12 to 1
Dad	0	0	0	0	0	0
His	0.6344	0.9524	1 to 1.5	0.2637	1.1986	1 to 4.55
Sister	0.0264	0	Girls only	0	0	0
Brother	0.0529	0	Girls only	0.024	0	Girls only
Him	0.185	0.2484	1 to 1.34	0.1918	0.2568	1 to 1.34
Princess	0	0	0	0	0	0
Girl	0.0529	0	Girls only	0.048	0	Girls only
Boy	0.0264	0.1656	1 to 6.21	0.0959	0	Girls only
Man	0.2379	0.0828	2.87 to 1	0.0959	0.5565	1 to 5.80
Grandma	0	0	0	0	0	0

Gender word comparison

Fig:19

Conclusions:

- The first twelve words in the 2018 folklores are the same as in the 1930's Dúchas folklores with the exception of one, but they in are in a slightly different order. These words must be important for writers in both 2018 and in the late 1930's.
- There is a high frequency use of food and drink words in the past compared to the present (Fig. 10) We also noticed a high frequency of farming related words that are no longer mentioned in present day folklores. (Fig.11.). Another pattern which emerged was the high use of weather words. (Fig.12) Perhaps the most interesting of all was the number of nature words in the old folklores that were not used in 2018. These observation might reflect how life has changed in Kentstown in the last 80 years. The village is now considered a commuter village and farming would no longer be the main occupation. It may also show that students spend less time out in Nature.
- In the 2018 folklores, there is a high frequency of words that do not appear in the 1930' Dúchas folklores . These included:

➢ School related words (Fig.4) Has school become more important and enjoyable in the students lives in 2018?

➢Formal titles (Viscount, Lord, Lady etc. Fig.7)

➢ Words from Generic folklores that they may have learned in school. i.e Fionn Mac Cumhaill .(Fig.5)

➢Words about themselves such as I, am and me. (Fig.8)

➢Contracted forms of not. (Fig.6)

➢Words to describe a person's job (Profession) Fig.6

Gender

- Anne Bayetto's gendered word use list showed that in her study boys and girls most often write about their own Gender. Our Findings agree with Bayetto's for certain words only. (Fig.19)

➢2018 data – Girl's only use the word 'girl' and boys used the word boy 6.27 times more than girls did.

➢1930's data- Girl's only use the word girl but girl's only mention the word boy as well. Interestingly the ratio of boys writing about men compared to girls is high (5.8:1) and boy's use the word 'he' more than girls (4.55:1)

- In our 2018 folklores girls refer to home, house, family & school with a higher frequency than boys (Fig.15) Bayetto found the same pattern. Boys used fish, river , Hills and Estate at higher frequencies than the girls (Fig.16) These are noticeably more outdoor and environmental words.
- The girls from the 1930's wrote about eggs, cream, churn and butter at higher frequencies than the boys (Fig.18) Perhaps this reflects their chores !

Lost Words.

- In the 1930's folklores they use 26 words that are removed from the Junior Oxford Dictionary in 2009 but the 2018 folklores only use 8 words that have been removed (Fig.3). Our results show that all the nature words are lost from the new folklores except for the word mussel. This highlights the need to use more nature words in Primary schools and the importance of publications like "The lost words" (1).

Readability.

There is a significant difference between the Flesch Kinkaid readability value for the Dúchas 1930's folklore (Average 6.7) and the 2018 folklores (Average 9.4) ,these were compared using a t-Test. p= 0.009, rejecting the null hypothesis.