To ish or not to ish?

Daphne Theijssen, Hans van Halteren, Tip Boonpiyapat, Anna Lohfink, Bas Ruiter, and Hans Westerbeek

Department of Linguistics, Radboud University Nijmegen

Abstract

In English, new adjectives can be coined by adding the suffix -ish. For instance, one can describe someone who acts like Arnold Schwarzenegger as *Schwarzeneggerish*. This paper investigates how the use of -ish is influenced by text characteristics (genre, formality) and author characteristics (gender, age). We used two corpora, the British National Corpus and the Blog Authorship Corpus. From our analyses of variance (ANOVAs) and logistic regression models, we learned that for the use of -ish it is probably more important what type of text you are writing than who you are. We also concluded that this type of research is seriously hampered by the absence of the kind of metadata needed for our type of research.

1 Introduction

Languages tend to be reasonably complete in the sense that they provide words for the things that people are likely to say to each other on a day-to-day basis. However, for those cases where one would need to say something not yet catered for, languages also provide means to form new words. Say one would like to express a quality for which there is no word, but the quality is well-represented by describing it as being like another quality or entity, e.g. an uncompromising, pushy and even violent attitude could be described as being *like Arnold Schwarzenegger*. This likening to existing words is fairly common, in fact so common that the English language provides the adjective-forming suffix -ish for it, so that the above description can be replaced by the single word *Schwarzeneggerish*.

In this paper, we investigate the use of this suffix as witnessed in corpus material.[1] We do not limit ourselves to really productive uses, but also include those words which have been created long ago and have managed to conquer a place in the dictionary. To be exact, we are interested in the representation of qualities by likening to something else (by use of -ish) and it does not matter if the actual formation of the -ish word was done by the current author or by, say, Shakespeare. We do limit ourselves to adjectives which have been formed by suffixation with -ish, excluding for instance words like *finish*, which does not contain the suffix -ish, but is a verb ending in *ish*. We also limit ourselves to those instances in which -ish bears the sense 'representing a likeness' (so not, e.g., *Finnish*).

[1] The research was carried out in the course 'Corpus-based Methods' in the research master Language and Communication, a collaboration between Radboud University Nijmegen and Tilburg University.

Proceedings of the 20th Meeting of Computational Linguistics in the Netherlands
Edited by: Eline Westerhout, Thomas Markus, and Paola Monachesi.

Our investigation focuses on the extralinguistic circumstances under which -ish words are being used. There are at least two aspects here that we want to consider. The first is the type of text in which the words are found. The Longman grammar of spoken and written English (Biber et al. 1999), e.g., tells us that we come across the phenomenon mostly in conversation and fiction, i.e. texts showing a more informal style. For corpora with metadata that have been designed with linguistic research in mind, such as the BNC (BNC Consortium 2007), we can use the genre classification to see whether this is indeed true.

The second aspect is the type of author that wrote the text. It has been remarked that the use of -ish is much more extensive among younger people. For instance, there is a Facebook fan page called "Adding *ish* onto the end of a word when describing something" with over 800,000 fans. And one can also imagine that a tendency for description by likening might be connected to the psychological makeup of the author and that there might be differences in use by men and women. Again, this can be investigated on the basis of corpus data, provided that the relevant metadata are available.

In summary, the research question addressed in this paper is the following: can the amount of use in a text of words formed with the suffix -ish, in the sense of a likeness, be shown to be dependent on a) genre or formality of the text and b) gender and/or age of the author of the text?

In the remainder of the paper, we first present related work (Section 2). Then we describe in detail which -ish words we have chosen to consider in our investigation (Section 3). In Section 4, we describe the collection of the data set on which we perform two statistical analyses. Section 5 shows our findings when applying analysis of variance (ANOVA), while in Section 6, we use logistic regression. A discussion is provided in Section 7, our main conclusion in Section 8.

2 Related work

A large part of the research on the suffix -ish has focussed on the methodology for measuring the productivity of suffixes in general (e.g. Nishimoto 2004, Plag 2006). The focus has mostly been on comparing different methods for measuring the productivity, and comparing the productivity of different suffixes. Plag (2006) showed that in the BNC, the range of different words to which -ish is added is not very large, compared to other suffixes like -ness and -ion. But relatively many were hapaxes, indicating that -ish is very suitable for the coinage of new words. Baayen (1994) showed that there is a clear relation between the productivity of various affixes and text type (genre). For the suffix -ly, for instance, he concluded that the text type and the author's individual preferences can overrule restrictions on the possible stems. The suffix -ish was not included in this paper.

Some researchers have focussed on the possible stems to occur with -ish. Byrd et al. (1986) argued that nouns with the suffix -ish tend to be short. They found no four-syllable word to occur with -ish. Spencer (2005) stated that -ish can be attached to whole phrases, e.g. *a why-does-it-have-to-be-me-ish expression*. Such instances are most likely to occur with very frequent and easily recognisable

phrases, he argued, but he was unaware of studies focussing on this type of suffixing. Prcic (1999) gave a nice overview of the possible stems preceding -ish, e.g. adjectives, nouns and numerals. He also discussed the dropping of a stem-final *e* (e.g. *blue - bluish*) and the doubling of a stem-final single consonant (*snob - snobbish*).
The combination of different suffixes, including -ish, has also been researched. The focus has either been on the order in which they can or cannot be attached (e.g. Hay and Plag 2004), their separability from the stem (e.g. Hay 2002), or their repetitive use (e.g. *boyishishness* in Plag and Baayen 2009). There is also some research on the history of the suffix -ish, for example in the works by Shakespeare (Neuhaus and Spevack 1975). The use of -ish to modify colours has also been the topic of study by a number of researchers (e.g. Moroney 2003, Rao and Lohse 1996). And there is also quite some work on children's acquisition of suffixes (e.g. Klibanoff and Waxman 2000).

Despite the fact that a lot of research has been carried out on the use of the suffix -ish, we are not aware of studies that examine the influence of characteristics of the speaker or writer (age, gender) and text (genre, formality) on the tendency to use -ish. The goal of the current paper is thus to provide such a study.

3 An inventory of *-ish* words

To investigate whether the use of -ish in the sense of likeness depends on extralinguistic characteristics, we needed an inventory of words containing the suffix in this sense. This section shows how we constructed this inventory and how we divided the words into different classes (treated separately in the statistical analyses).

3.1 Corpora and type extraction

Since the targeted suffix is productive, it was not possible to base our inventory on existing lexicons. Instead, we had to extract all potential instances from the corpora we planned to use. At the start of our investigation, we selected five corpora, listed in Table 9.1. Even though it later turned out that only BNC (BNC Consortium 2007) and BAC (Schler et al. 2006) were usable for analysis, we based the inventory on all five in order to keep the list as general as possible.

Corpus	Number of words
British National Corpus (BNC)	100,000,000
Wikipedia XML Corpus 2006 (WIKI)	350,000,000
Blog Authorship Corpus (BAC)	140,000,000
Caroline Tagg's Txt Msg Corpus (CorTxT)	189,000
Web-as-Corpus kool ynitiative UK (ukWaC)	2,000,000,000
Total	2,590,000,000

Table 9.1: Size of corpora at the basis of the inventory

We extracted all words from these corpora by splitting on white space, and kept those which ended in *ish* (ignoring word-final non-alphanumeric characters). Out of the 2.6 billion words present in the corpora, we extracted 4,025,285 tokens showing a total of 20,199 types.

3.2 Filtration

However, not all these 20,199 types were instances we want to consider. We automatically removed all URLs (477 types, especially frequent in the ukWaC corpus) with regular expressions specifying that the to-be-removed type contained strings like `www`, `http` and `.co.uk`. We also excluded the words (44 types) that consisted of four or fewer letters (e.g. *ish*, *yish*), making an exception for instance where the first character is a number (e.g. *6ish*). Finally, we used CELEX (Baayen et al. 1993) to remove the 210 types which were verbs (e.g. *finish*), common nouns (e.g. *parish*) or adjectives with an initial capital letter (e.g. *English*). The number of types we removed in this step may seem relatively small (731 of 20,199), but they account for over 90% (3,633,639) of the 4,025,285 tokens in the five corpora.

3.3 Manual classification

After this step of automatic removal, there were still 19,468 types left. We performed a pilot study to check the feasibility of manual identification of those types that can be split into a base and -ish, and where -ish does represent likeness. Three annotators, who were all non-native but fluent speakers of English, judged the relevance of the 1000 most frequent -ish words in the list of 19,468. The Kappa scores reached between the different pairs of annotators were low: -0.16, -0.04 and 0.44. We therefore had to conclude that the task proves too difficult.

The pilot study also made clear why this task was so difficult. There were many clear cases with recognisable stems, e.g. *foolish*, *greenish* and *silly-little-me-late-again-ish*. But to recognise other instances, one needed a vocabulary of a size that would have been daunting to many a native speaker, e.g. *priggish* and *marish*. This was also true for instances like *skittish*, which should not be included since -ish here had another sense than likeness. The need for vocabulary knowledge turned into a need for world knowledge when the type in question was capitalised: one needed to know that there is a singer called Björk to be able to approve *Bjorkish*, or that *Hammish* is just a less frequent spelling of the name *Hamish*.

3.4 Automatic classification

Having determined that manual annotation would lead to inconsistent, not well-analysable data, we decided to fall back on more trustworthy resources. We only kept those -ish words that on the basis of these resources could be classified as numeral-based (e.g. *6pmish*), name-based (e.g. *Beatles-ish*), noun-based (e.g. *bunnyish*) or adjective/adverb-based (e.g. *happy-ish*). A disadvantage of this procedure is that almost 80% (16,056) of the types could not be classified into the four classes. We thus fail to identify some of the more creative type such as

hahahahahahahahahahahahahahahahahahaha-ish. There is a risk that these creative types are exactly those characteristic for the language of younger people. However, this disadvantage is outweighed by the advantage of replicability.[2]

Numeral-based -ish words

In order to recognise numerals with the suffix -ish, we constructed two lexicons based on the findings in the corpora. The first was the *number lexicon* which consisted of cardinal numbers written out, e.g. *twelve* and *million*. Moreover, irregular ordinal numbers (e.g. *first* and *second*) and words related to times (e.g. *o'clock*, *quarter* and *p.m.*) were included. The second lexicon was the *unit lexicon*, consisting of units that we encountered in our data, e.g. *year-old*, *miles* and *hrs*. An -ish word was added to the numeral category if: (1) the first alphanumeric character of the word was a digit, and the stem consisted only of digits, non-alphanumeric characters and/or units in the unit lexicon (e.g. *13miles-ish*), or (2) the stem contained no capitals and consisted only of digits, non-alphanumeric characters, the coordinator *and* and/or numbers in the number lexicon (e.g. *two-thousandish*). Of the 20,199 types, 958 were found to have a numeral as base.

Name-based -ish words

For names, a different strategy was needed, as no comprehensive list of names could be built. We decided to include only the most certain types, being those capitalised forms with explicit marking of -ish, e.g. *Eeyore-ish* or *Jackass(ish)*. Types that occurred in the corpora in such a marked form could also be accepted when occurring without explicit marking. For such instances, the ratio between the number of occurrences with and without explicit marking should be below 50 in the five corpora. This for instance excluded *Kurdish*, which occurred 13,438 times, but only 1 time as *Kurd-ish*, and included *Schwarzeneggerish*, since this form and *Schwarzenegger-ish* both occurred once. Moreover, when changed to lower case, the stem should not be an adjective (e.g. excluding *Newish*), a common noun (e.g. excluding *Bullish*) or a numeral (e.g. excluding *Threeish*). This procedure led to a list of 1,001 types that are name-based.

Noun-based and adjective/adverb-based -ish words

For -ish applied to nouns, adjectives and adverbs, we only included those instances for which the base was present as such in CELEX (Baayen et al. 1993). For the base, we did allow inflected forms such as plural nouns and comparative/superlative adjectives and adverbs. Following Prcic (1999), we derived three potential bases from each -ish type: the actual form (e.g. *green* from *greenish*), the actual form with an *e* (e.g. *blue* from *bluish*) and the actual form but undoubling any double final consonant (e.g. *red* from *reddish*). We applied this procedure to all types without capitals and with at least one vowel.

[2]Our -ish inventory is available at `http://lands.let.ru.nl/~daphne/downloads`.

There were some bases which were ambiguous as to class, e.g. *chocolate*, which could be interpreted as a noun and an adjective/adverb according to CELEX. We disambiguated these manually: We preferred noun for stems that were intuitively hard to interpret as an adjective (e.g. *fool*) and that were materials (e.g. *copper*). The rest, including colours (e.g. *blue*, *blonde*) and directions (e.g. *west*, *up*), we preferred to label as adjectives/adverbs.
After dividing all these -ish words into the categories *noun-based* and *adjective/adverb-based*, we manually checked all noun-based types which had at least 5 instances in the corpora. We removed those words where -ish showed another sense than likeness, e.g. *stylish*, forms that were probably spelling errors, e.g. *finsish*, and uncapitalised names, e.g. *hamish*. The final set of noun-based -ish words contained 1,476 types, and that of adjectives/adverbs-based -ish words consisted of 708 types.

4 Collection of a data set

Having established four lists of words with the suffix -ish with the desired meaning, the next step was to prepare the corpus data for the analysis of the degree of their presence. For this we did not only need to count the selected words in each text, but also had to determine for the text the extralinguistic features we wanted to relate the degree of use of -ish words to: the genre/formality of the text, and the gender and age of the author. This section mainly describes how we determined these features, but returns at the end to the degree of use of -ish words.

4.1 Availability of metadata

As stated above, we employed five corpora to establish our inventory of -ish words (WIKI, CorTxT, BNC, BAC and ukWaC). However, only BNC and BAC provide the desired author information at the text level. The other three corpora do not contain metadata at all, Wikipedia and ukWaC being automatically crawled and lacking such data completely and CorTxT having been anonymised. The genre is provided in the BNC, but not for BAC. For the latter, we had to determine genre/formality in some way on the basis of the text itself.

For the statistical analyses, we thus limited ourselves to two corpora: BNC and BAC. They differ on a number of levels: Where the BNC contains only British English, the type of English in BAC is unknown. The BNC contains language from various sources, while the BAC contains only blogs from `blogger.com`. Most of the data in the BNC stems from the period 1985-1995, while most of the data in the BAC comes from 2000-2004. Because of these differences, we treated the use of -ish in the corpora separately in this paper.

4.2 Gender and age

For the BAC, all 19,320 texts are marked with the gender and age features. This is not the case for the BNC: although the documentation states that gender and age are present, we found that this is true for only 603 of the 4,050 texts. A case in

point is the spoken part of the corpus (908 texts), for which no speaker information is present. For instance, for the demographic part, it turned out that it is not the gender and age of the speakers that is available, but rather the gender and age of the respondent, the person who is walking around with the recorder. For the written part, 234 texts could not be used because they are marked as having mixed authorship and 2,304 because the author's age and/or gender is not documented. The author's age is classified into six classes in the BNC: 0-14, 15-24, 25-34, 35-44, 45-59 and 60+. The first of these, 0-14, contains very few texts, so we decided to merge it with the class 15-24, leaving us with five classes spanning all possible ages. The author ages in the BAC fall into three classes: 13-17, 23-27 and 33-48. Authors with ages outside these three ranges were not included in the corpus, in order to have clearly separable groups for the corpus' main goal: authorship recognition and profiling. Since we analysed the two corpora separately, we decided to use the two classification schemes as is and not to attempt to merge them.

4.3 Genre and formality

Genre

The genre feature is only present for the BNC. The full genre scheme in BNC's metadata distinguishes no fewer than 70 different genres. In order to have sufficiently large classes for analysis we merged the genres of the 603 remaining texts into four major classes: academic (101 texts), non-academic (115 texts), fiction (244 texts) and the rest (143 texts). The largest subgenres in the rest category were *misc* (57 texts), *biography* (49 texts), *religion* (13 texts) and *commerce* (12 texts).

Formality

As already mentioned, we expect genre to be mostly effective because the genres are linked to a level of formality. Since only the BNC contains information about genre, we needed a measure for the formality of the text, to replace genre when necessary. We used the Flesch reading ease score (Flesch 1948) as an approximation of the formality. This measure, together with other estimations of sentence complexity, is very commonly used in automatic genre classification (see for instance the overview in Luštrek 2007). The Flesch reading ease score is a number that usually varies between 0 and 100 (but it can be even higher or lower), and for which a higher score is found to correlate with simpler text. It is designed to go down when a text contains more words per sentence (longer sentences), and more syllables per word (longer words). To be exact, it is calculated as follows:

$$reading\ ease = 206.835 - 84.6 * (y/w) - 1.015 * (w/s)\ ,$$

where y is the number of syllables in the text, w is the number of words in the text and s is the number of sentences in the text.

We used GNU Style and Diction[3] to calculate the reading ease score for each

[3]See `http://www.gnu.org/software/diction/diction.html`.

text. We then used this score to establish the formality, being a binary feature which had the value *Informal* when the reading ease score was over a threshold and *Formal* otherwise. Based on the measurements for BNC (fiction having mean 85.1 and standard deviation 6.3 and academic plus nonacademic having mean 53.5 and standard deviation 9.8), we set the threshold at 73.0.
For the BNC, 97.5% of the academic and non-academic texts were considered *formal* in this measure, and 96.8% of the works of fiction *informal*. In the rest category, 68.7% of the texts were labelled *formal* and 31.3% *informal*.

4.4 Final selection

For the BAC we could use all 19,320 texts. For the BNC, as already mentioned, we removed all texts with insufficient metadata, all spoken texts and all texts with mixed authorship. Furthermore, we limited ourselves to recent English, removing texts from before 1985 (54 texts). The final set comprised 549 out of the 4,050 texts. Table 9.2 shows the average use of -ish words in the two sets.

	BNC			BAC		
Type of base	N	mean	stdev	N	mean	stdev
adjectives/adverbs	549	25.6	75.0	19,320	16.2	122.6
names	549	0.3	2.7	19,320	2.4	43.1
nouns	549	69.5	134.0	19,320	49.3	214.1
numerals	549	1.7	20.6	19,320	15.7	125.7

Table 9.2: Number of texts (N), mean number of -ish words per million (mean) and standard deviation (stdev) for the different categories and corpora.

The standard deviations show that there is a lot of variation between the texts. Moreover, we see that noun-based -ish words are the most common. Considering the four categories as one category (disregarding the categorisation) could yield results that are mostly based on the noun-based -ish words. To avoid this, we kept the four categories apart in the statistical analyses. Because the use of numeral-based and name-based -ish words is very small in the BNC, we excluded the analysis of these two categories for the BNC. For the BAC, we kept all four categories.

5 Analysis of variance (ANOVA)

We used analysis of variance (ANOVA) to test whether the average value of the dependent variable (the use of -ish per million words) significantly differs between certain groups, taking into account individual variation within the groups.

5.1 Method

We employed a factorial analysis of variance (ANOVA) with three fixed factors (age, gender and genre/formality).[4] Only the interaction between gender and age was included. Since the dependent variable needs to have a distribution that is (close to) normal, we excluded all instances without an -ish word (i.e. having value *0*) and for the remaining instances took the log of the number of instances per million.
In ANOVA, the log of the use of -ish per million is calculated as follows:

$$X_{ijkl} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \gamma_k + U_{ijkl} \ ,$$

in which X is the use of -ish per million for individual text l with a writer aged i and having gender j, having genre/formality k. The symbol μ represents the overall average, α_i the average deviance from μ in this age group, β_j the average deviance from μ in this gender group, $\alpha\beta_{ij}$ the average deviance from μ in this age and gender group, γ_k the average deviance from μ in this genre/formality group, and U_{ijkl} the individual deviance from μ. In order to establish the effect of the factors, ANOVA calculates the total sum of squares of the deviances from the average μ.

5.2 Results

The ANOVAs for noun-based and adjective/adverb-based -ish words in the BNC can be found in Tables 9.3 and 9.4. They show that there are no significant effects for the -ish words with an adjective or adverb as base. For the noun-based -ish words, gender, genre and formality have a significant effect on the use of -ish per million words. When we look at the average values, we see that females use noun-based -ish words much more frequently than men: on average, females use 131 noun-based -ish words per million (in the texts that contained at least one occurrence), males only 96 per million. With respect to genre, the presence of -ish is smallest in academic and nonacademic texts (61 per million), and largest in fiction (141 per million). The same effect is found for formality: formal texts have relatively few -ish words per million (80 per million), and informal texts relatively many (128 per million). For both adjectives and nouns, however, there is still a lot of variance unexplained, seeing the large sums of squares (SS) for the residuals.
The ANOVAs for the use of numeral-based, name-based, noun-based and adjective/adverb-based -ish words in the BAC set can be found in Tables 9.5 and 9.6. The formality of the text only has a significant effect on the use of numeral-based and adjective/adverb-based -ish words. It has the same pattern as the noun-based -ish words in the BNC: on average, formal texts show fewer occurrences of -ish words (237 per million for numerals, 195 per million for adjectives/adverbs) than informal texts (283 per million for numerals, 230 per million for adjectives/adverbs). Moreover, we find a (near) significant effect of age range

[4] We used the function `aov()` in the `stats` package in `R`.

	Noun				Adj/Adv		
	df	SS	F		df	SS	F
Gender	1	12.3	20.0	***	1	0.4	0.7
AgeRange	4	4.2	1.7		4	1.0	0.4
Genre	3	18.3	9.9	***	3	1.3	0.7
Gender:AgeRange	4	2.7	1.1		3	1.5	0.8
Residuals	342	210.9			166	105.9	

Table 9.3: ANOVA for noun-based and adjective/adverb-based use of -ish in BNC texts, using *genre*, *** p<0.001 ** p<0.01 * p<0.05 · p<0.10.

	Noun				Adj/Adv		
	df	SS	F		df	SS	F
Gender	1	12.3	19.1	***	1	0.4	0.7
AgeRange	4	4.2	1.6		4	1.0	0.4
Formality	1	6.0	9.3	**	1	0.1	0.1
Gender:AgeRange	4	4.3	1.7		3	1.7	0.9
Residuals	344	221.5			168	107.0	

Table 9.4: ANOVA for noun-based and adjective/adverb-based use of -ish in BNC texts, using *formality*, *** p<0.001 ** p<0.01 * p<0.05 · p<0.10.

for all types in this corpus, whereas there were no significant effects of age range found in the BNC data. The youngest bloggers, aged 13 to 17, use -ish words most frequently for all base categories except names. For name-based -ish words, the oldest bloggers, aged 33 to 48, are the most frequent users, although the difference is only near significance. As with the BNC, the high residuals show that a large part of the variance remains unexplained.

	Numeral				Name			
	df	SS	F		df	SS	F	
Gender	1	0.4	0.2		1	0.2	0.1	
AgeRange	2	138.8	41.5	***	2	11.2	3.0	·
Formality	1	13.9	8.3	**	1	1.6	0.9	
Gender:AgeRange	2	4.1	1.2		2	3.1	0.8	
Residuals	1079	1804.9			285	530.8		

Table 9.5: ANOVA for numeral-based and name-based use of -ish in BAC texts, *** p<0.001 ** p<0.01 * p<0.05 · p<0.10.

	Noun				Adj/Adv			
	df	SS	F		df	SS	F	
Gender	1	0.0	0.0		1	0.1	0.0	
AgeRange	2	85.6	36.6	***	2	78.8	25.7	***
Formality	1	0.8	0.7		1	7.6	5.0	*
Gender:AgeRange	2	4.6	2.0		2	2.5	0.8	
Residuals	3333	3901.7			1384	2125.8		

Table 9.6: ANOVA for noun-based and adjective/adverb-based use of -ish in BAC texts, *** p<0.001 ** p<0.01 * p<0.05 · p<0.10.

6 Logistic regression

We also addressed our research question by considering it as a classification problem: we used a modelling or machine learning technique to predict whether the number of -ish words in the text is above-average or not. So, using the per-million-counts for each text in our BNC data set, we established the average number of noun-based and adjective/adverb-based -ish words per million, and for each text checked whether they were above average or not. The same was done for the BAC data, also for the name-based and numeral-based -ish words.

Because the texts in the BAC were relatively short (7,004 words on average), over 76% (14,789) of the texts in the BAC did not contain an -ish word at all. The texts in the BNC set were over five times longer on average (35,716 words), and consequently only 29% (159) of the texts lacked an occurrence of -ish. As a result, there was a higher imbalance between the above and below average use of -ish in the BAC set than in the BNC set.

We experimented with a number of techniques, including Naive Bayes, Ripper, C4.5 and Logistic regression. The first three, being classifiers, suffered from the imbalance in the BAC. The high majority baseline (86% to 98%) made it difficult for the algorithms to learn patterns that were able to improve this baseline. Logistic regression is much more robust with respect to class imbalance (Owen 2007). We thus limit ourselves to logistic regression in this paper.

6.1 Method

We modelled the use of -ish as depending on the values assigned to the same features as in the previous section: the age of the writer, his/her gender, the interaction between the two, and the genre or formality of the text.

In logistic regression modelling, a regression function is established that fits the data matrix best. It yields the log of the odds that the use of -ish in the text in question (I) is 1 (above average) rather than 0 (below average):

$$\texttt{logit}[p(I = 1)] = \alpha + \beta X$$

where α is the model intercept, X are the feature values and β are the coefficients. The coefficients β can be understood as the weights assigned to the features by the model, where positive values increase, and negative values decrease, the odds that there is above average use of -ish. The optimal values for α and β were estimated using Maximum Likelihood Estimation[5].

The log odds were used to establish the quality of the model. In this paper, we use the area under the ROC curve (AUC), which gives the probability that the regression function, when randomly selecting a positive (above average use of -ish) and a negative (below average use of -ish) instance, outputs a higher log odds for the positive instance than for the negative instance.[6]

6.2 Results

The AUC reached by the models built on the BNC data can be found in the top two rows of Table 9.7. Although their values are not really high, they are sufficient to allow an analysis of the coefficients in the model, presented in Table 9.8. The results are similar to what we found for the ANOVAs in Section 5. Writers are more likely to use a noun-based or an adjective/adverb-based -ish word when writing a work of fiction (the coefficients 0.93 and 0.57 are both positive), and less likely when writing an academic text (-1.33 and -1.41). Informal texts are more likely to show above average use of -ish (1.11 and 1.04) than formal texts.

	Numeral	Name	Noun	Adj/Adv
BNC model with genre			0.751	0.687
BNC model with formality			0.708	0.658
BAC model with formality	0.592	0.547	0.536	0.561

Table 9.7: AUC reached by the regression models

The AUCs for the four regression models built on the BAC data are presented in the bottom row of Table 9.7. The scores are very low; the models are only slightly better than chance (AUC=0.5). The regression coefficients are presented nonetheless in Table 9.9. As in the ANOVAs, there are many more significant effects than in the BNC. The formality of the text is significant for numeral-based and adjective/adverb-based -ish words: informal texts tend to have more above average use of -ish for these categories than formal texts. There are also some significant effects of gender and age. For age range 13-17, we see that the use of noun-based -ish words is likely to be lower than for the other age ranges. This is the opposite of what we saw in the previous section, where the average use of noun-based -ish words was highest for the youngest bloggers. This difference, the small values of the coefficients and the low AUC show that it may be unadvisable to draw strong conclusions from the coefficients in this model.

[5] We used the function `glm()` in the `stats` package in `R`.

[6] We used the function `somers2()` created in `R` by Frank Harrell.

Feature	Noun		Adj/Adv	
(Intercept)	-0.68 **	-1.08 ***	-0.70 **	-1.23 ***
Female	0.34	0.35	0.02	0.04
Age 0-24	-14.56	-14.91	-1.02	-1.29
Age 25-34	0.12	0.18	-0.74	-0.70
Age 45-59	-0.23	-0.36	-0.04	-0.14
Age 60+	-0.13	-0.22	0.34	0.29
Academic	-1.33 ***		-1.41 ***	
Fiction	0.93 ***		0.57 *	
Nonacademic	-0.63 *		-0.61 *	
Informal		1.11 ***		1.04 ***
Female, Age 0-24	15.38	15.30	-13.71	-13.72
Female, Age 25-34	0.14	-0.10	1.10	0.88
Female, Age 45-59	0.08	0.31	-0.08	0.03
Female, Age 60+	0.44	0.59	-0.35	-0.29

Table 9.8: Coefficients of BNC regression models,*** p<0.001 ** p<0.01 * p<0.05 · p<0.10.

Feature	Numeral	Name	Noun	Adj/Adv
(Intercept)	-3.73 ***	-4.01 ***	-1.72 ***	-3.11 ***
Female	0.40 ***	0.00	0.15 *	0.43 ***
Age 13-17	0.21 *	-0.36 *	-0.20 **	0.08
Age 33-48	-0.37 *	-0.42	-0.13	-0.02
Informal	0.68 ***	0.04	-0.03	0.24 **
Female, Age 13-17	-0.09	-0.02	-0.05	-0.07
Female, Age 33-48	-0.14	0.11	-0.20	-0.05

Table 9.9: Coefficients of BAC regression models,*** p<0.001 ** p<0.01 * p<0.05 · p<0.10.

7 Discussion

For the BNC, we have been able to confirm the finding put forward in the literature that formality is a major factor in the amount of use of -ish. Both statistical analyses showed that -ish is applied to noun bases significantly less in more formal texts than in more informal texts. This effect could be shown both on the basis of the genre classification provided in the metadata and on the basis of an automatically derived substitute, our formality measure based on the Flesch reading ease score. For adjective/adverb-based -ish words, we only found an effect of genre and formality in the logistic regression models, but it showed the same pattern. As to the gender and age of the author, only ANOVA discovered a significant influence on the application of -ish in the BNC, but only for those with noun bases. It seems that females more frequently use noun-based -ish than males, but this is not confirmed by the logistic regression model. It could be that our inventory of -ish words is too limited to discover more (clear) differences between age and gender groups, since

the more creative types have been left out of consideration.

For the BAC, only the application of -ish to numeral and adjective/adverb bases could be shown to be significantly influenced by formality with both methods. For the age and gender of the writer, it appears there are significant influences. There seems to be a significant effect of age range on the use of -ish for all four bases, although it is not confirmed by the logistic regression model for adjectives/adverbs. Moreover, some coefficients in the logistic regression model showed different effects than the average values used for the ANOVAs. It seems that these observations are not unequivocal, as both methods reported low model quality (by residuals and AUC). Apparently, the factors we included in the analyses are not enough to explain the variance in the data; there is still much individual variation left.

It may even be that the influences we found are only indirect, being an artifact of specific gender/age groups writing about specific topics, which we could not take into account due to the absence of metadata. We have thus observed (yet again) that for an investigation into linguistic variation, one not only needs large amounts of textual data, but comprehensive metadata as well. The currently available corpora tend to fail in at least one of these demands. The BNC, although in principle well-supplied with metadata, in the end provides relatively few texts for which the desired metadata is present, potentially too few to be able to show significance of the relevant influences. The BAC, although containing larger amounts of text, only provides very limited metadata, while larger corpora, such as ukWaC, lack the kind of metadata that are needed for linguistic research altogether. Here, factorial statistical analysis becomes impossible from the start.

8 Conclusion

In this paper, we investigated whether the amount of use in a text of the English suffix -ish, in the sense of likeness, is dependent on genre/formality of the text and the gender and age of the author. We did this by applying two types of statistical analysis, ANOVA and logistic regression, to two corpora, a selection of written texts from the British National Corpus (BNC) and the Blog Authorship Corpus (BAC). The large unexplained variance of the ANOVA (the residuals) and the low confidence of the logistic regression models (the AUC) showed that we could only partly explain the use of -ish. The results seem to indicate that, as to the use of -ish, it is probably more important what you are writing than who you are.

For our study of the suffix -ish, we have been partially able to circumvent the lack of metadata (at least for the BAC) by replacing the genre/formality feature by an automatically derivable substitute based on the Flesch reading ease score. But this was only one feature, representing only one dimension of the in principle much richer genre feature. Given the potential of the extraction of huge corpora from the internet, but their lack of metadata, we think that a good way forward would be to attempt to induce some classification similar to the metadata in BNC automatically from the text. Although classifying gender and age appear a bridge too far at the moment, features like genre and domain ought to be feasible. Of course, automatic induction of such metadata has many possible drawbacks and

pitfalls. Still, we believe that such automatic induction is a fruitful avenue of future research. Once we manage to generate these metadata automatically, there will be a basis for a much larger number of investigations into linguistic variation than is possible now.

References

Baayen, R. Harald (1994), Derivational productivity and text typology, *Journal of Quantitative Linguistics* **1** (1), pp. 16–34, Routledge.

Baayen, R. Harald, Richard Piepenbrock, and Hedderik van Rijn (1993), *The CELEX Lexical Database (CD-ROM)*, Linguistic Data Consortium, University of Pennsylvania, Philadelphia, USA.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan (1999), *Longman grammar of spoken and written English*, MIT Press.

BNC Consortium (2007), *The British National Corpus, version 3 (BNC XML Edition)*, Oxford University Computing Services. http://www.natcorp.ox.ac.uk/.

Byrd, Roy J., Judith L. Klavans, Mark Aronoff, and Frank Anshen (1986), Computer methods for morphological analysis, *Proceedings of the 24th annual meeting of the Association for Computational Linguistics*, pp. 120–127.

Flesch, Rudolph (1948), A new readability yardstick, *Journal of applied psychology* **32** (3), pp. 221–233.

Hay, Jennifer (2002), From speech perception to morphology: Affix ordering revisited, *Language* **78** (3), pp. 527–555, Linguistic Society of America.

Hay, Jennifer and Ingo Plag (2004), What constrains possible suffix combinations? On the interaction of grammatical and processing restrictions in derivational morphology, *Natural Language & Linguistic Theory* **22** (3), pp. 565–596, Springer.

Klibanoff, Raquel S. and Sandra R. Waxman (2000), Basic level object categories support the acquisition of novel adjectives: Evidence from preschool-aged children, *Child Development* **71** (3), pp. 649–659, Blackwell Publishers.

Luštrek, Mitja (2007), Overview of automatic genre identification, *Technical Report IJS-DP-9735*, Jožef Stefan Institute, Department of Intelligent Systems, Jamova 39, 1000 Ljubljana, Slovenia.

Moroney, Nathan (2003), Unconstrained web-based color naming experiment, *Proceedings of the SPIE*, pp. 36–46.

Neuhaus, H. Joachim and Marvin Spevack (1975), A Shakespeare dictionary (SHAD): Some preliminaries for a semantic description, *Computers and the Humanities* **9** (6), pp. 263–270, Springer.

Nishimoto, Eiji (2004), Defining new words in corpus data: Productivity of English suffixes in the British National Corpus, *26th Annual Meeting of the Cognitive Science Society (CogSci 2004)*.

Owen, Art B. (2007), Infinitely imbalanced logistic regression, *The Journal of Machine Learning Research* **8**, pp. 773, MIT Press.

Plag, Ingo (2006), Productivity, *Handbook of English Linguistics* pp. 537–556.

Plag, Ingo and R. Harald Baayen (2009), Suffix ordering and morphological processing, *Language* **85** (1), pp. 109–152, Linguistic Society of America.

Prcic, T. (1999), The treatment of affixes in the 'big four' EFL dictionaries, *International Journal of Lexicography* **12** (4), pp. 263, Oxford University Press.

Rao, A. Ravishankar and Gerald L. Lohse (1996), Towards a texture naming system: Identifying relevant dimensions of texture, *Vision Research* **36** (11), pp. 1649–1669, Elsevier.

Schler, Jonathan, Moshe Koppel, Shlomo Argamon, and James Pennebaker (2006), Effects of age and gender on blogging, *AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*.

Spencer, Andrew (2005), Word-formation and syntax, *Handbook of word-formation*, Springer, pp. 73–97.