A corpus-based investigation of junk emails

Constantin Orasan and Ramesh Krishnamurthy

Computational Linguistics Group
School of Humanities, Sciences and Social Sciences
University of Wolverhampton
Stafford Street, Wolverhampton, WV1 1SB
United Kingdom
C.Orasan@wlv.ac.uk, ramesh@easynet.co.uk

Abstract

Almost everyone who has an email account receives from time to time unwanted emails. These emails can be jokes from friends or commercial product offers from unknown people. In this paper we focus on these unwanted messages which try to promote a product or service, or to offer some "hot" business opportunities. These messages are called junk emails. Several methods to filter junk emails were proposed, but none considers the linguistic characteristics of junk emails. In this paper, we investigate the linguistic features of a corpus of junk emails, and try to decide if they constitute a distinct genre. Our corpus of junk emails was build from the messages received by the authors over a period of time. Initially, the corpus consisted of 1563, but after eliminating the duplications automatically we kept only 673 files, totalising just over 373,000 tokens. In order to decide if the junk emails constitute a different genre, a comparison with a corpus of leaflets extracted from BNC and with the whole BNC corpus is carried out. Several characteristics at the lexical and grammatical levels were identified.

1. Introduction

Emails have become a normal part of life for many people, who send and receive them on a daily basis. One of the main advantages of using emails is that they allow almost instantaneous communication between people on different sides of the world. Unlike telephone calls, emails also provide a record of information exchanged and decisions made. Another major advantage of emails is that they are usually cheaper than other communication methods.

Unfortunately, there are people who exploit this ease of electronic communication by sending large quantities of unwanted emails, trying to convince people to buy products and services, to invest their money in various financial opportunities and business ventures, or to protest against political oppression or environmentally-unfriendly activities that are taking place in different parts of the word.

The general term used for unsolicited emails is *spam*. In this paper we deal only with *junk emails*, a subcategory of *spam* which includes mainly commercial emails. In section 2, we discuss in more detail what we consider to be a junk email.

In some cases the people who send us spam emails are friends, but in most cases these messages are sent by people who are completely unknown to us, often from email addresses which do not actually exist, or which are abandoned after a few weeks. We emphasise this aspect, because the receiver ideally wants to have most of these messages removed automatically, and this cannot be done just by removing all the messages received from a certain address.

One alternative would be to filter out all the messages coming from a certain domain, but such a method would have disastrous effects given that most of the junk emails come from popular services like HotMail and Yahoo. Banning messages from these domain addresses would filter out many legitimate messages from friends and colleagues.

Junk emails are not a new phenomenon, nor an entirely unexpected one, as they were predicted as early as 1975 in the Internet Request For Comments (Postel, 1975), where it was noted that the transmission protocol does not include a feature to refuse messages from a certain host. Unfortunately, as mentioned earlier, with the widespread availability of free email services, such a measure against junk email is no longer appropriate.

There have been attempts to pass legislation to ban the use of emails to promote services and products, but this has proved to be only partially successful. Extensive information about lawsuits, news, and opinions about junk emails can be found at http://www.junkemail.com and http://spam.abuse.net/. However, a discussion of the legal status and litigational aspects of junk emails is beyond the scope of this paper.

Due to the partial failure of legislative bodies to impose laws which ban the use of email for sending unrequested offers, computer system administrators started to look into the possibility of filtering out junk emails automatically. The big challenge posed by such filters is that they obviously must not filter out any legitimate messages, but they have to be efficient enough not to let too many junk emails pass through. Some of these filters are rule based, involving human beings observing junk emails and writing rules which can filter them out. Given the amount of work required to design such rules by hand, machine learning methods have also been used (Sahami et al, 1998; Androutsopoulos et al, 2000; Carreras and Marquez, 2001). The results reported in these papers are very impressive, but all these methods did not look at the messages as such, but fed them into a machine learning algorithm, letting the algorithm decide how to recognize a junk email.

In this paper, we try to identify junk emails from a linguistic perspective attempting to establish the characteristics of junk emails as a genre. It should be emphasized that the goal of this paper is not actually to produce a new junk email filter.

Given that emails have come into widespread use relatively recently, they have understandably not yet been

investigated thoroughly by corpus linguists and computational linguists. As a result of this, it is not really known what distinguishes the junk emails genre from other genres. In this paper we try to find some of the characteristics of junk emails by comparing them with the BNC, a general language corpus, (Burnard, 1995) and with a corpus of leaflets extracted from BNC. The comparison with the BNC is to see how different the junk emails are from a general corpus. The comparison with the corpus of leaflets is to see if it is possible to say that junk emails are an electronic equivalent of leaflets.

This paper is structured as follows: in section 2 we define what a junk email is. The corpus of junk emails built and used in this research is described in section 3. One of the main problems we had to face while building the corpus was to find a way to remove duplications. The automatic method devised to remove duplications is presented in section 4. The analysis of the junk emails corpus and the investigation of the characteristics of the junk emails genre are described in sections 5. We finish the article with our conclusions.

2. What is a junk email?

Deciding what is a junk email is difficult. In this paper we define a *junk email* as an email that is received without being requested, and which tries to promote a product or service, or to offer some "hot" business opportunity. Examples include emails which try to sell "miracle" products for losing weight, try to persuade the reader to join some quick self-enrichment scheme, or suggest shares which will apparently increase their price rapidly.

However, we do not suggest that every email promoting a product is a junk email. For example, people frequently receive promotional messages as a result of subscribing to a mailing list which belongs to a company. It may of course be the case that the receivers are not aware that they have subscribed to the mailing list, because the information is often contained within text that is written in a very small font (as is also the case with many leaflets). A major difference between junk emails and emails received from a company is that you can unsubscribe from the mailing list of a company, whereas with junk emails this is not necessary possible (even though in many junk emails, the option is specifically offered).[1]

We consider junk emails to be a type of spam email, but we are not trying to investigate spam emails, because there is an even greater problem in deciding whether an email is spam or not. In general it is accepted that in addition to junk emails, other types of spam emails are chain letters, jokes, letters asking us to join a protest campaign, etc. The main difference between junk emails and the other types of spam emails is that in most cases these other spam messages are sent by friends who think that the recipient might be interested in the text content. Because these messages usually come from a friend, they should not be filtered out.

Deciding whether an email is a junk email or not is becoming increasingly difficult. Although many of messages are straightforward offers for products and services, an increasing number of junk emails appear to address the recipient in a more personal manner, making it more difficult to decide whether the email is unsolicited or not. One feature which seems to be increasingly used is the name of the recipient in the salutation. In addition to this, the junk emails sound more and more like emails from friends in their style and tone, e.g. "Dear Jay, Here is the information you requested.", "Hi, …. Best regards, Happy Jonny".

A few tips to spot junk emails are offered at http://www.junkemail.org/scamspam/fraud.shtml:

- Messages offering business opportunities which promise large rewards for very little effort and risk
- Investment opportunities usually in shares which are supposed to increase their price rapidly
- Credit repair schemes and credit card offers with extremely low interest rates
- Miracle health products, especially for losing weight or for improving your sex life
- Travel offers at a very attractive prices

We have enumerated these tips because when we analyse our corpus of junk emails, features indicating any of these categories will be identified.

One of the most interesting features of junk emails is how quickly they can adapt to a new political and economical situation, and give a solution to it. After the Anthrax cases in the United States, we noticed an increasing number of messages offering remedies (including natural remedies) for Anthrax, and an increasing number of offers to buy antibiotics from Mexico.

3. The corpus

In order to carry our investigation, a corpus was built. Over a period of time we saved all the junk emails we received, because we realized the necessity of having such a corpus. Building a large corpus of junk emails is not a trivial task. First of all, most of the emails are quite short, so that it is difficult to assemble large amounts of text. Secondly, not knowing how to find a place where junk emails are stored and therefore could be downloaded, the only way to build such a corpus is to store the ones we ourselves received. During the phase of collection, we noticed that there is not a wide variety of junk emails sent in the same period of time. As we show in section 4, many of the emails received are duplications of ones we already have, so the only solution would have been to postpone this research and wait till the corpus had grown much larger.

The decision to classify an email as junk email was taken by us using the definition provided earlier. In the initial stages of the building of the corpus, we did not know how to eliminate the duplications from our corpus. For this reason, whenever we spotted a duplication, the message was removed. This means that the number of messages which we included in our corpus was much

[1] The general belief is that if someone uses the option to unsubscribe from a junk email mailing list, he or she will receive more junk emails, because he/she is confirming that there is a genuine individual reading the emails received at that email address. We cannot confirm or refute the validity of this belief. However, while building the junk emails corpus, the first author of this paper tried unsubscribing in order to collect more junk emails, but none of the addresses offered for unsubscription requests was valid!

smaller than the actual number we received. However, in our linguistic research all these duplicated messages do not tell us anything about the genre of junk emails and therefore they had to be eliminated.

At present, the entire corpus of junk emails contains 1563 files, which amount to around 880.000 tokens. However, it should be emphasised that this corpus contains duplication, which were removed by the automatic method presented in the next section. In order to facilitate processing, the corpus was encoded in XML. An example of an encoded message is the following: [2]

```
<JUNKEMAIL>
<SUBJ>$79,000 a month?</SUBJ>
<DATE>Wed, 16 Jan 2002 16:09:20</DATE>
<FROM>"Andrew West" webmaster@korcin.com</FROM>
<MSG>
Please, please allow me to take your precious
time just for a second!

Do you want to make $79,000 a month?
Do you want to retire in March 2002?
Do you want to tour around the world with your
family?
Do you want to take part in events with your
family around the world?
Do you want freedom of time and money from March
2002?

If yes, send back blank mail. If not, send back
with "REMOVE".

This is the program for the common people to be
rich.

Thanks.

Andrew
</MSG>
</JUNKEMAIL>
```

As can be seen in the example, the subject is marked with **<SUBJ> … </SUBJ>**, the date when it was sent is marked with **<DATE>…</DATE>,** and the sender with **<FROM>… </FROM>**. We thought it was worthwhile to keep all this information because, even though it is not directly relevant for this paper, it can be used by other researchers for different purposes. For example, the date when the message was sent can be used to find out whether there is any link between the political and economical situation and the type of junk emails sent (e.g. as we mentioned earlier, the case of Anthrax). The sender can be used to find out whether certain domains are preferred by senders of junk emails. The actual text of the message is enclosed between the **<MSG>** and **</MSG>** tags.

When this corpus was being built, two types of messages were considered: plain text messages and HTML encoded messages. We also noticed other types of messages containing different types of attachments such as RTF files, images, and even applications. All these messages were completely ignored. The plain text messages were easy to process by a simple perl script. For those messages encoded in HTML, the *lynx* program from Linux was used to extract the text. All the images and other non-textual information were ignored. Unfortunately, by ignoring the images, some information was lost, because an increasing number of junk emails contain references to images which are downloaded automatically when the email is opened, and which are meant to make the message more attractive.

4. The problem of duplications

One of the common problems which has to be addressed when building a corpus is to avoid duplication. In other types of corpora, duplication can be easily avoided by carefully selecting the sources of the textual material. When a corpus of junk emails is built, such an approach cannot be taken. This is because we had a passive role in the acquisition of the messages. We did not request them, they were sent to us over a long period of time. For this reason, a method to identify duplications was required.

Different methods to identify duplication have been used and documented, but none seemed appropriate for us. For example, when building the Bank of English (www.cobuild.collins.co.uk/boe_info.html), one method of checking large numbers of text files for duplicate texts was as follows:

a) the Unix *sort* program was used to sort all the text lines from the files into alphabetical order
b) the Unix *uniq* program was used to identify duplicated text lines
c) the duplicated lines were manually checked in the original text files, to see whether the texts themselves were completely identical, or merely contained some identical lines (e.g. the same quote by a politician used in two different newspapers)
d) any identical texts were removed

We could not use this method for the corpus of junk emails, because in the end it requires that the files which are thought to be identical have to be manually checked. This method could not be used in our case because of the large number of very similar messages. Another reason why this method could not be employed is that it requires that the files are identical in order to signal this fact. Because formatting is used in junk emails, two messages saying the same thing are not necessary identical.

Any method which relies on the sender's name or the subject was also out of the question because, as mentioned earlier, copies of the same junk emails are very frequently sent with the sender's name changed, or the subject changed. In fact, as everyone who has received junk emails has probably realized, the subject is not a very good indicator of the actual content of a junk email, as they are often very general or completely irrelevant (e.g. "Hi", "great news", "I thought you might be interested").

The method that we had to design needed to take into account any small differences between messages, and given the large number of comparisons we had to perform, needed to be fast. After experimenting with different formulas, we decided that a good measure to compute the difference between two messages is given by the following formula:

[2] The whole corpus can be downloaded for free for research purposes from http://clg.wlv.ac.uk/projects/junk-email/

$$Diff = \frac{remove + insert}{length_1 + length_2}$$

Where:

- *remove* and *insert* are the number of words to be removed and inserted in the first file in order to obtain the second one
- $length_1$ and $length_2$ are the length in words of the first and the second message, respectively

The values for *remove* and *insert* are computed using the *diff* command from Linux, and $length_1$ and $length_2$ are computed using the *wc* command. For this reason the value for *diff* is computed very fast. This formula was chosen because it leads to high values when different messages are compared, because a large number of words have to be inserted and removed in order to obtain the same file[3]

After using this method, we noticed that the *diff* values between different messages tend to cluster around 0.9, whereas for very similar files they cluster around 0.1. Random checking revealed that the measure indicates correctly which messages are similar and which are not. After observing the values for *diff* we decided to consider any messages identical if they had a *diff* value of less than 0.3. When two or more files were identified as being identical the largest one was kept for our corpus. After filtering all the duplications from our corpus, its size decreased dramatically. If at the beginning our corpus had 880.000 tokens, after filtering the number of tokens dropped to 373.000 tokens, in 673 files. This means that almost two thirds of the messages were duplications.

After designing our algorithm to eliminate duplications, we found a paper proposing a similar, but slightly more complex method (Broder, 1998).

5. Comparison with other corpora

In order to identify the characteristics of the junk email genre, we compared it with a corpus of leaflets extracted from BNC, to see if the junk emails are merely an electronic version of the leaflets. To make the comparison easier, we made the size of the corpus of leaflets comparative with the size of junk emails corpus. We also compared the junk emails and the leaflets with the entire BNC corpus, to measure them against general language.

5.1. Sentence length

In general, the sentences in the junk emails corpus are fairly short, 48% are less than 10 words, compared with 27% for leaflets and BNC. The average sentence length was 15.81 words for junk emails, 19.09 for leaflets and 21.66 for BNC. However, sentence boundaries are difficult to establish, with the additional problems caused by formatting and encoding, so we will not rely too heavily on this metric.

[3] The extreme situation is when the files are completely different, and therefore $length_1$ words have to be removed and $length_2$ have to be inserted, making the value of the *diff* measure 1.

5.2. Part of speech distribution

All three corpora were tagged using the FDG tagger (Tapanainen and Jarvinen, 1997). The distribution of the tags by types and tokens is given in Tables 1 and 2 below.

POS tag	Freq of POS-tag in 10000 types		
	Junk	Leaflets	BNC
N	3902.77	5135.87	6065.72
A	843.68	1516.50	1365.86
V	683.97	1011.00	342.61
NUM	577.51	476.32	1043.51
ABBR	446.82	191.23	528.87
EN	237.41	440.82	157.16
ADV	234.89	366.30	104.89
ING	234.89	428.87	221.93
PREP	60.80	77.34	26.77
PRON	53.99	88.23	10.15
DET	35.07	46.75	14.87
CS	15.39	22.50	2.52
CC	7.32	11.25	9.58
INFMARK	3.03	4.57	0.46
NEG-PART	1.26	1.76	0.44
OTHER	2661.22	180.69	104.65

Table 1: Part Of Speech distribution: by types

The leaflets corpus seems to have the highest frequency of types for each part of speech, followed by junk emails and then BNC. This suggests that the leaflets have a wider vocabulary, which may be due to the fact that they cover a greater range of topics. This confirms our initial casual observations that junk emails are restricted to a small number of products and services. Junk emails have the lowest type frequency for nouns and adjectives, which suggests more repetition.

The frequency of untagged types (OTHER) in junk emails is largest by several orders of magnitude. This may be explained by the number of unusual words in junk emails (e.g. *kiff*, *aphrodisia*), but also has the effect of lowering the frequency values for the other tags.

POS tag	Freq of POS-tag in 10000 tokens		
	Junk	Leaflets	BNC
N	1859.55	2406.36	2076.95
V	809.67	1046.59	1105.31
PRON	573.35	591.67	725.39
PREP	555.98	939.25	922.02
DET	474.51	817.32	843.07
ABBR	418.34	46.43	72.22
A	398.51	675.17	580.16
ADV	329.31	401.53	494.05
CC	327.72	320.57	285.49
NUM	282.22	192.67	181.26
EN	111.66	217.71	226.16
INFMARK	105.19	141.32	130.41
ING	91.62	131.01	134.01
CS	73.25	102.97	121.98
NEG-PART	43.50	37.96	63.64
OTHER	3545.60	1931.48	2037.88

Table 2: Part Of Speech distribution: by number of tokens

When considering tokens, the distinction between junk emails and leaflets is clearer. The values for leaflets are closer to the BNC. There are one or two exceptions: junk emails and leaflets are more similar with respect to pronouns and conjunctions. The high frequency of abbreviations in junk emails may be a feature of the electronic medium, as well as reflecting the variety of organizations involved. The low frequency of determiners and adjectives suggests that they are often omitted in junk emails, in order to make them shorter and more direct.

5.3. Lexical frequencies

5.3.1. N-grams

One way of analysing the lexical contents of the corpus is by creating *ngram* frequency lists. This collects consecutive sequences of *n* number of lemmatised tokens, and shows us the most common sentences or part-sentences. Frequency lists for 20-grams to 2-grams were generated and analysed. In this section a few selected features are discussed.

In the BNC, the most frequent 20-gram was `take hold of your top leg by the ankle and pull back the leg as far as possible without straining` (16 occurrences), followed by `whilst his firm were the auditors of a limited company issued an audit report on that company 's accounts for` (13). Both of these are so topic-specific that they suggest a lot of duplication in the corpus.

In the leaflets, no 20-grams occurred more than twice! In the junk emails, however, the most frequent lexical 20-grams (i.e. ignoring sequences of characters used for formatting such as ***** and %%%%%) were `f r e e f r e e f r e e f r e e f r e e` (24) and `please click below and enter you email at the bottom of the page you may then rest-assured that you will` (22). The emphasis on *free* is typical of this genre, despite the influence of formatting in this case. The second 20-gram is a recurring message at the end of junk emails, ostensibly promising a way of preventing further junk emails from the same sender (but, as we stated in Section 2, it is unlikely to be effective!).

Some of the emails go to great lengths to deny their junk status: `under bill s 1618 title iii pass by the 105th US congress this letter can not be consider spam`. There is clearly some awareness of existing anti-junk-email legislation and its stipulations.

Further down the frequency list, we get reassurances about the efficacy of the procedures involved, their legality, and their profitability: `you ship option I would like to receive I package fedex overnight I be include $15 for ship Hawaii &` (5), `with all software once open the cd may not be return however if find defective it will be replace with` (4) and `this have help to show people that this be a simple harmless and fun way to make some extra money` (4).

Now let us look at shorter ngrams. The 9-grams list for BNC is similar to the 20-gram list, in that it again suggests duplicated documents: `net profit for the year to december 31 was` (44), `the general entry requirements for admission to a first` (43) and `silver anniversary couples receive a bottle of sparkling wine` (42). However in the leaflets, similar repetitions are not duplicated texts, but fixed paragraphs within texts: `accident data total casualty in this age group be` (7), `Parcelforce national enquiry centre on 0800 22 44 66` (6) and `all you have to do be complete the attach` (5).

In the junk emails, 9-grams give more indications of the actual products and services: `one 2.0 oz jigget bar of Kathmandu temple kiff` (15), `to become a millionaire utilize the power of multilevel` (6), `Great opportunity to make relative easy money with little` (6) and `one 1 oz bottle of sweet vjestika aphrodisia drop` (6).

Some emails request postal addresses, fax numbers, telephone numbers, credit card details, etc.: `to order to order by phone call 530-343 9681` (6), but a few give very strange instructions: `make sure the cash be conceal by wrap it` (7).

Junk	Leaflets	BNC
254 to *be remove*	171 if **you** *be*	17398 one of the
226 **you** *do* not	114 one of the	9855 the end of
195 *be remove* from	111 *be able* to	9682 as well as
183 in the subject	102 there *be* no	8279 I *do* n't
166 if **you** *do*	92 part of the	8105 part of the
145 if **you** *be*	91 **you** *will* find	7819 there *is* a
139 on the internet	89 there *be* a	7479 some of the
130 **you** *will be*	86 the regional council	7478 out of the
115 the subject line	85 **you** *will be*	6602 a number of
114 *remove* in the	84 **you** *do* not	6592 end of the
110 if **you** *have*	77 if **you** *have*	6222 it *was* a
105 *would like* to	71 to *help* **you**	6060 there *is* no
102 *remove* from we	68 the number of	6020 the fact that
99 one of the	64 the end of	6008 there *was* a
91 **you** *want* to	61 it *be* a	5889 *be able* to
86 **you** *would* like	59 a number of	5645 to *be* a
85 *click* here to	58 *have to be*	5511 in order to
83 if **you** *would*	52 to *ensure* that	5478 it *is* not
82 *be able* to	51 **you** *want* to	5400 per cent of

Table 3: Comparison of 3-gram frequency lists

Comparing the three corpora, we see that junk emails have a very distinct set of 3-grams, whereas there are more matches between leaflets and BNC. Common phrases in BNC and leaflets decrease in junk emails (e.g. *one of the*, *part of the*, *a number of*). There are few topic-specific items in BNC, but in leaflets we see *the regional council*.

In junk emails, we see *on the internet*, but also the even more specific phrases relating to the junk email genre: *to be remove*, *be remove from*, *in the subject*, *the subject line*, *click here to*, *you email address*, and *the link below*.

In fact, we find three distinct sets of sequences in the junk emails:

a) sequences referring to the prevention of future junk emails: `to be remove, be remove from, in the subject, remove in the`

b) sequences specific to online communications: `on the internet, the subject line, click here to, you email address, the link below`

c) sequences recognizable from general language: `you do not, if you do, if you be, you will be, if you have`

One striking feature is the prominence of the pronoun *you* (representing the direct appeal by the messages to the recipients; highlighted in bold in the list). Another

immediately observable feature is the prominence of verbs (highlighted in italics in the list).

5.3.2. Lemma frequencies

The prominence of *you* in junk emails from the 3-gram list (8 of the top 20 items, compared to 6 in leaflets and 0 in BNC) is confirmed when we look at lemma frequencies. *We* is similarly prominent in junk emails, but *I*, *it* and other pronouns are more prominent in the other corpora. However, verbs (which were prominent in junk emails 3-grams) are less evident in the lemma lists (even *be* and *have* are used less than in leaflets and BNC). This is probably another indication of the terse, telegrammatic style of junk emails, as also evidenced by the much greater use of & rather than *and*. The reduced occurrence of *of* in junk emails also indicates the use of shorter noun groups.

		Frequency per 10000 lemmas		
Lemma	POS	Junk	Leaflets	BNC
You	PRON	35.64	28.53	11.66
The	DET	31.13	55.43	84.88
Be	V	25.73	36.43	57.40
&	CC	21.31	0.67	0.26
And	CC	21.09	28.52	36.65
Of	PREP	16.33	28.74	40.66
To	INFMARK	16.04	17.32	22.02
A	DET	14.85	20.75	29.02
We	PRON	13.51	8.68	7.26
For	PREP	10.74	13.15	11.48
In	PREP	10.44	16.18	25.70
To	PREP	9.62	10.93	13.44
I	PRON	9.51	1.65	16.21
Have	V	7.96	8.87	18.88
This	DET	7.45	4.20	5.09
It	PRON	7.15	8.90	17.42
Not	NEG-PART	6.69	4.71	10.89
Will	V	6.61	7.45	4.62

Table 4: Comparison of Lemma frequencies

Table 4 showed the commonest lemmas in the corpora, whereas Table 5 presents a few lemmas selected because of their known association with the junk email genre (e.g. from the www.junkemail.com list).

Word	Junk	Leaflets	BNC
Risk	0.30	0.39	0.21
Free	3.48	1.19	0.33
Money	2.06	1.15	0.53
Business	2.52	1.18	0.55
Investment	0.52	0.42	0.17
Credit	1.68	0.59	0.12
Quick	0.27	0.27	0.27
Fast	0.45	0.23	0.16
Internet	1.64	0.00	0.00
Email	3.84	0.00	0.00
e-mail	2.66	0.01	0.00
Sex	0.28	0.03	0.12
Weight	0.37	0.26	0.14
Miracle	0.05	0.00	0.02

Table 5: Normalised lemma frequencies (per 10,000)

With the exception of *risk*, all the items are more commonly used in junk emails than in leaflets or BNC. This confirms the topic orientation of the majority of junk emails (i.e. business opportunities, sex aids, weight-loss and other "miracle" products, and internet business services). In the case of *risk*, it is mainly in combination with *free* that it occurs in junk emails. It is very interesting that whereas *quick* is evenly distributed between the corpora, *fast* is preferred in junk emails.

Email addresses and web addresses are common in junk emails, whereas they were mostly absent from leaflets and BNC.

6. Conclusions

After some initial discussion, we defined a junk email as an email that is received without being requested, and which tries to promote a product or service, or to offer some "hot" business opportunity. A further feature is that with a genuine junk email, it is difficult or impossible for the recipient to prevent the receipt of further junk emails from the same sender. As senders become more sophisticated, for example by adopting the salutations, style and tone of interpersonal emails, it is becoming even more difficult to distinguish junk emails from other emails.

We collected junk emails we received ourselves over a long period of time in order to build up a corpus large enough to be worth analysing and comparing with other corpora. Initially the corpus contained 1,563 messages totalising over 880,000 tokens, but we noticed that this corpus contained duplications.

The process of eliminating duplicates was our next concern, and we suggested a fairly simple but effective methodology for doing so. After applying the automatic method to remove duplications the size of the corpus reduced to about 373,000 tokens in 673 messages.

Finally, we compared our junk emails corpus with a corpus of leaflets (representing a similar but older genre, printed rather than electronic) extracted from the BNC, and with the whole BNC corpus (representing the language in general).

Junk emails are written in shorter sentences than leaflets or other BNC texts. However, the problem of

identifying sentence boundaries led us to be cautious about this measure.

The leaflets had a higher type frequency for almost every part of speech, suggesting that they contain a wider range of vocabulary than junk emails. Junk emails had fewer nouns and adjectives, suggesting more repetition. Junk emails contain more untagged types, probably because they contain more unusual words.

Leaflets and the BNC were closer in token values. The values for junk emails were lower, except for pronouns, conjunctions, and abbreviations. Determiners and adjectives are particularly low in junk emails, again suggesting compression and shorter noun groups.

The lexical analysis first made use of *ngrams*. The 20-gram frequency lists showed up duplication in the BNC. Leaflets contained no recurring 20-grams, but the 20-grams in junk emails highlighted *free* offers, junk-status denials and (useless) unsubscription advice. 9-gram lists again showed duplicated texts in the BNC, as opposed to standardized, formulaic, or legally obligatory paragraphs in leaflets. In junk emails, 9-grams started to reveal more specific information about the types of products and services being offered.

The 3-gram lists indicated that junk emails had a very distinct phraseology, whereas leaflets and BNC were more similar to each other. Junk emails lack many of the common phrases of the general language (e.g. *one of the, part of the, a number of*), but feature two other sequences instead: prevention of future junk emails, and general online phrases.

Lemma frequency lists confirmed the prominence of the pronoun *you* in junk emails, which had been noticed in the 3-gram lists. The reduced occurrence of verbs (and determiners and adjectives, as seen in the earlier POS analysis) indicates a very terse and economic style. Selecting a few typical lemmas from the junk emails genre, we saw a completely consistent view of the distribution of lemmas across the corpora: junk emails highlighted *free, money, investment, credit, fast, internet, email, sex, weight*, and *miracle*!

7. References

Ion Androutsopoulos, John Koutsias, Konstantinos Chandrinos, Constantine D. Spyropoulos, 2000 An experimental comparison of naive bayesian and keyword-based anti-spam filtering with personal e-mail messages. In Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, (SIGIR2000), July 24-28, Athens, Greece, pp. 160-167

Andrei Z. Broder, 1998 On the resemblance and containment of documents. In *Compression and Complexity of Sequences (SEQUENCES'97)*. pp. 21 – 29, IEEE Computer Society

Lou Burnard, 1995 *Users Reference Guide: British National Corpus Version 1.0*, Oxford University Computing Services, UK.

Xavier Carreras and Lluis Marquez, 2001: Boosting trees for anti-spam email filtering, In *Proceedings of RANLP2001*, Tzigov Chark, Bulgaria, pp. 58 - 64

J. Postel, 1975 On the junk mail problem. Network working Group Request for Comments: 706, NIC #33861, November, http://www.faqs.org/rfcs/rfc706.html

M. Sahami, S. Dumais, D. Heckerman and E. Horvitz (1998) A Bayesian Approach to Filterin Junk E-amis. In Learning for Text Categorisation – Papers from the AAAI Workshop, pp. 55 – 62, Madison Wisconsin. AAAI Technical Report WS-98-05

John M. Sinclair 2001, Preface. In Ghadessy, M., Henry, A. and Roseberry, R. L. (eds) *Small Corpus Studies and ELT: Theory and Practice*, John Benjamins

P. Tapanainen and T. Jarvinen, 1997 A non-projective dependency parser. In *Proceedings of the 5th Conference of Applied Natural Language Processing*, pp. 64 – 71, Washington D.C., USA