Assessing algorithms for automatic extraction of anglicisms in Norwegian texts

Dr Gisle Andersen
Department of Language, Culture and Information Technology (Aksis)
University of Bergen
gisle.andersen@aksis.uib.no

1 Introduction

English loan words thrive in many languages, including Norwegian. Anglicisms may be used for different purposes, to denote new concepts (such as *reality tv* in [1]), as new words for known concepts (such as *party* in [2]), in movie titles [3], brand names [4], names of companies, products, artists, and so on.

[1] Folk følger **reality tv** på CNN. God underholdning. (DB030324)
People are watching reality TV on CNN. Good entertainment.

[2] Men vi hadde **party** i går, og holdt prøveavstemning. (BT050529)
But we had a party last night, and held a test vote.

[3] At ”**Kingdom of Heaven**” har en balansert framstilling av korstogene høres lovende ut. (DA050429)
That ”Kingdom of Heaven” gives a balanced presentation of the crusades sounds promising.

[4] Jeg har en **Smartphone** med **Windows** allerede, og tror jeg skal kjøpe en nyere modell nå, sier Falch. (AP050523)
I already have a Smartphone with Windows, and I think I'm going to buy a newer model now, says Falch.

The use of recent loan words raises debate and criticism, and the Norwegian Language Council encourages the formation of domestic words to replace recent loans in cases like [1]-[2]. The critics rarely express the same antagonism towards loan words of other origins, like *sudoku* and *focaccia*, or towards older, well established loan words of English origin that are no longer perceived as such, like *sport* and *pledd* (plaid). Setting aside this debate, the current paper is concerned with how anglicisms can be identified in Norwegian texts, and it describes the development of a language processing tool that is used to automatise the identification process.

The paper reports on a series of experiments involving alternative strategies for automatic anglicism retrieval: n-gram-based pattern matching and lexical database lookup, as well as combinatory methods. The tool's performance is evaluated against manually controlled data. The n-gram statistics are mainly gathered from the BNC via chargram lists produced by the “Phrases in English” project (http://pie.usna.edu/explorec.html), but restricted to productive and uniquely English chargrams, in the sense that they are non-existent in a Norwegian reference lexicon. The lexicon lookups are mainly based on existing data developed in the SCARRIE project (http://www.ling.uib.no/~desmedt/scarrie/).

2 Material: The Norwegian Newspaper Corpus

The processing tool described in the current study is developed as part of the Norwegian Newspaper Corpus project (http://avis.uib.no). The Norwegian Newspaper Corpus is a large and self-expanding corpus of Norwegian newspaper texts (Hofland 2000). The collection of this dynamic and continually growing corpus began in 1998. The corpus is automatically updated by means of *w3mir* (http://langfeldt.net/w3mir/), which is an all-purpose http copying and mirroring tool. On a daily basis, the mirroring tool retrieves recently published texts from a set of remote web sites, specifically the entire Internet version of ten major Norwegian newspapers. A set of own-developed tools is used for further processing and annotation of the texts. The system automatically selects the relevant text, ignoring advertisements, navigation menus, metatext, html code and so on. Next, it automatically classifies the text as either *bokmål* or *nynorsk* – the two forms of Norwegian that exist, also identifying and rejecting texts that are entirely written in English. Further, the texts are annotated with word class and other morphosyntactic information by means of the Oslo-Bergen tagger, and the tagged and untagged texts are added to the database.

Approximately 200,000-250,000 running words are added per day. As of June 2005, the database consists of about 370 million words, and it is by far the greatest searchable corpus of Norwegian.

The selection of newspapers that are included allows for comparison across various categories: broadsheet versus tabloid formats, national versus regional newspapers, and general content versus business and finance newspapers. The corpus is accessible via a web interface that uses IMS Corpus Workbench (http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/). A new project web site with more detailed information and corpus statistics was launched earlier in 2005.

In addition to the text collection and annotation procedure briefly described, the set of corpus tools daily creates a list of newly encountered word forms. Each word form of all the new texts is compared against a comprehensive list of known word forms. This list is a compilation that includes all word forms gathered in connection with various corpus and lexicographical projects carried out at Aksis over the years, including the full form lexicon *Bokmålsordboka* and the lexical resources developed in connection with the Oslo-Bergen Corpus and tagger (http://decentius.aksis.uib.no/cl/cgp/obt.html). It consists currently of about 3.2 million unique word forms. The new word forms that are not found in the long reference list are then stored in an archive of neologisms. This archive naturally provides a good resource for the study of word formation processes, lexical productivity, linguistic creativity and so on, and it is considered a valuable resource among Norwegian lexicographers (Wangensteen 2002).

A language processing tool automatically classifies the latest neologisms according to morphologically distinct categories such as hyphenated or non-hyphenated compounds, names, digits, acronyms, name-lexeme combinations, and so on, and the latest newcomers in each category can be viewed on the project web site. The current paper describes the module for identification of anglicisms, which is an integral part of this classification tool.

The neologism archive is large, and comprises roughly 127,000 items as of June 2005. On average, about 1,500 new word forms are encountered every day. Preliminary studies of the material (including Andersen 2004a, 2004b) indicate that the most common types of neologisms found in this material are compounds, names (including acronyms), spelling errors, and, indeed, anglicisms. Compounds are particularly prevalent, and appear to account for 40-50 per cent of the neologisms. This is crucially linked to the fact that in Norwegian, like the other North Germanic languages and German (but unlike English and French), compounds are as a rule written as one word, without a hyphen (although hyphenation is optional). Consequently, an overt result of journalistic linguistic creativity is the emergence of content-rich, new compounds, often found in newspaper headlines.

The current study is based on a 10,000-word subset of this neologism archive, described more in detail in section 4.1. These words were manually and automatically classified, and the experimentation amounts to comparison of alternative classification methods with the 'golden standard' of manually classified words.

3 Anglicisms in Norwegian

3.1 Why anglicisms?

The identification of anglicisms is potentially relevant for linguistic, language-political and language-technological purposes, and the existence of a computational tool to perform this task may be useful in order to reduce the need for manual work. Such a tool is obviously helpful as a means for retrieving new loan words among the vast amount of new words that emerge. The information will enhance lexicographical work, in that lexicographers will have more direct access to words that may be relevant for inclusion in dictionaries, as well as knowledge about their origin. Language-politically, the information that a word is an anglicism will be instrumental in assessing the overall foreign influence on Norwegian language and assessing whether a language such as English poses a real threat towards the existence of Norwegian as a separate language.[1]

Moreover, the detection of words of English origin is useful for language-technological purposes. In text-to-speech synthesis, recognising that a word is of foreign origin should lead to the activation of conversion rules from orthography to phonology that are different from those applied for domestic words. Such a tool may also be beneficial for term extraction purposes in the identification of new terminology in various domains. Also, the correct identification of anglicisms has an additional project-internal function, as it may be used for quality-assurance for the identification of English texts, which sometimes occur among the retrieved texts, but which are not meant for inclusion in the Norwegian Newspaper Corpus (cf. section 2).

As mentioned, the focus of the current paper is on loan words of English origin, but also loan words of other origins may be detected using the methods proposed below. By far, English outnumbers other languages as a source of loan words in Norwegian, but it appears that also Japanese and Italian loan words are common in Norwegian today. Loan words of other origins than English are excluded from the current study.

The aim of the study is to process words that have not been previously recorded, so the focus of the study is on **recent** anglicisms. This means that well-established loans words like *sport* and *pledd* (cf. section 1) are not relevant here and are not classified as anglicisms in the manual classification.

3.2 Compounding and 'norwegification'

Anglicisms may come in different forms, and typically they retain their English orthography when they are used in Norwegian contexts, as shown in [5] below.

[1] In fact, an officially appointed committee is currently engaged to assess this issue; cf. http://www.sprakrad.no/templates/Page.aspx?id=7450.

[5] Det viste seg at siden vi fikk så kort varsel så hadde vi ikke nok **crew** til å gjøre intervjuet. (DB990131)
It turned out later that we were given such a short notice that we did not have enough crew to do the interview.

[6] Stilsikkert i **croonertradisjonen** (DB050111)
True to the style of the crooner tradition

[7] Likevel var hun 67 sekunder bak Simone Luder (24) under det første **parkverdenscupløpet**. (AP020511)
Nevertheless, she was 67 seconds after Simone Luder in the first park world cup race.

[8] Raser mot **barnereality** (DB050613)
Raging against child reality

[9] Det var en skikkelig **døll** plass. (DB010810)
It was a really dull place.

[10] Det hjelper ikke enslige mødre eller arbeidsledig ungdom med kropper som blir vandaliserte av **jønkfood** og sinn som sultefores av voksen tafatthet. (SA030328)
It does not help single mothers or unemployed youth with bodies which are vandalised by junk food and minds which are starved by adult indolence.

[11] Siden Tolkien satt i kveldinga og **nerda** med alvespråkene sine, er verden blitt okkupert av … (SA011219)
After Tolkien sat in the evenings and [nerded] with his elfish languages, the world has been occupied by …

In addition to constituting words in their own right, it is very common that anglicisms are used as part of compounds, where the English component may appear in initial, medial or final position in a word, as shown in [6]-[8]. Some lexemes are clearly very productive as compound components.

Occasionally, English loan words get a normalised, or 'norwegified' spelling, which corresponds more closely to Norwegian pronunciation than the English orthography does. This is the case with the common adjective *døll* 'dull', which, in fact, grossly outnumbers the original English spelling *dull* in the corpus; cf. [9]. Sometimes the norwegification is only partial in compounds, as shown in *jønkfood* 'junk food' in [10]. However, there may be considerable variability in the spelling of anglicisms, and typically, alternative forms, like *chat – chatt*, *chating – chatting*, etc. co-occur.

As is well known, loan words may undergo various morphological, syntactic, and semantic processes after they are adopted in a new language. This may lead to new meanings or syntactic functions of a word that are non-existent in the source language. An evident example from the corpus is the use of *nerda* as a verb, as shown in [11]. This is a development of *nerd*, which has undergone conversion from noun to verb and has a Norwegian past tense ending *-a*. Also, an item like *shotsene* 'the shots' illustrates the common phenomenon that an originally plural form is borrowed, and combines with a Norwegian plural morpheme.

3.3 Challenges for automatic classification

Due to their normalised spelling, items like *døll* and *jønk* follow the normal pronunciation rules of Norwegian, and are therefore unproblematic with respect to conversion from orthographic to phonological representation in a speech-technological setting. For this reason, I have chosen to exclude anglicisms with a norwegified spelling from the current investigation.

English and Norwegian orthography are relatively disparate, which to some extent enables a rule-based automatic identification of anglicisms on the grounds of a word's orthography. For example, the consonants c, q, w, x and z are not found in domestic words but generally indicate a foreign origin. But a foreign orthography does not necessarily entail that a word has English origin, of course. Well-established loan words from other languages may also have foreign elements, such as *scene* from Greek. Conversely, because many anglicisms have an orthography that is in fact quite similar to Norwegian, it is clear that a strictly rule-based method may be problematic. A recent borrowing like *date* has no overt indication that it stems from English. Besides, English-Norwegian homographs cause problems for recent loans like (*web*) *hosting*, which has a Norwegian orthographically identical form (meaning 'coughing').

A retrieval method based solely on lexicon lookup may also be a challenge. As mentioned, many anglicisms are part of compounds consisting of a foreign and a domestic element; cf. *croonertradisjonen* in [6] above. Therefore a word-level lexicon lookup will not suffice, but coupled with morphological decomposition of a word, this method may be useful.

4 Tool for automatic classification of new words

As briefly mentioned, the algorithm for automatic extraction of anglicisms is a sub-process of a more comprehensive tool for automatic classification of new words (Andersen 2004a). The classification and identification tools are written in Perl. The anglicism module is a separate subroutine which takes as input an orthographic string and returns a value for a classification variable that indicates a word's status as an anglicism or not. At the moment, the tool operates on individual items at word-level, but an implementation of multiword unit processing is planned. It is also the intention to implement a finer classification of morphologically decomposed word forms where the anglicism classification applies to individual parts of words, such as compound components.

In fact, the anglicism tool contains five alternative subroutines performing the same classification task but using different methods, as described in section 4.2 below. In order to test the performance of the individual subroutines, I manually classified a subset of 10,000 forms as a reference against which the output of the tool was checked. A specially developed comparison tool, also written in Perl, produced both statistical output and lists of correct and incorrect automatic classifications.

4.1 Manual classification

The subset used for test data was selected randomly from the lists of neologisms dating from January to March 2005. Totally 72,319 new words were processed and randomised; 33,448 were selected as candidates for experimental testing. These were items that were not orthographically marked in any way, i.e. they did not containing a hyphen, digit or initial capital letter. However, the subset did contain many new compounds, loan words and some spelling errors. Of these, a randomised subset of 10,000 items was picked out for manual classification.

The actual classification was crude; it simply amounted to marking each word as either an anglicism or not. An item was considered an anglicism if it constituted an English word form in its own right, or if it contained at least one compound element of English origin. Occasionally, there may be several such elements, as in *popcroonerstemme* 'pop crooner voice'. If in doubt of a word's possible English origin, I consulted the *Anglisismeordboka* ('Dictionary of Anglicisms') by Graedler & Johansson (1997) as an authoritative source, or I checked the corpus occurrence of a word form in more detail.

As regards well established and integrated loan words, like *sport*, *robot*, *golf* and *mobbe* ('to bully'), I decided not to include them. I also excluded foreign items like *cell*, *pizza* and *nettjournal* 'net journal', which stem from other languages than English. A word like *megaviktig* 'very important', containing the common prefix *mega-*, did not count as an anglicism since the prefix is an earlier borrowing from Greek via Latin, although it is possible that recent English usage may have affected Norwegian users in this case. Controversially, perhaps, I did include the peculiar loan *fotballbager* 'football bags', containing the lexeme *bag*, which is originally from Old Norse but has been recently re-borrowed from English.

4.2 Alternative methods for automatic extraction of anglicisms

A series of five experiments were performed, involving alternative strategies for anglicism retrieval. The first two experiments use a lexicon lookup method, firstly by checking if a new word form resembles words that are found in the BNC, secondly by comparison against a tailored lexicon consisting of uniquely English forms (cf. sections 5.1-5.2 below). The third experiment (cf. 5.3) uses pattern matching based on a list of chargrams, that is, combinations of characters that are found to be typical in English (specifically, in words that occur in the BNC), but untypical in Norwegian, since they occur infrequently in a Norwegian reference lexicon. Next, a method using a set of manually constructed regular expressions was tested (cf. 5.4). The regular expressions were written on the basis of general knowledge of English orthography. Finally, the last experiment was a combinatory method using chargrams and regular expressions (cf. 5.5).

5 Experiments and results

In the test database containing 10,000 neologisms, 563 items were manually classified as anglicisms, amounting to an overall ratio of 5,63 per cent. Figure 1 below gives the overall results of the experimental testing, as briefly outlined in section 4.2. The results of each phase of the experimentation are discussed in greater detail in sections 5.1-5.5.

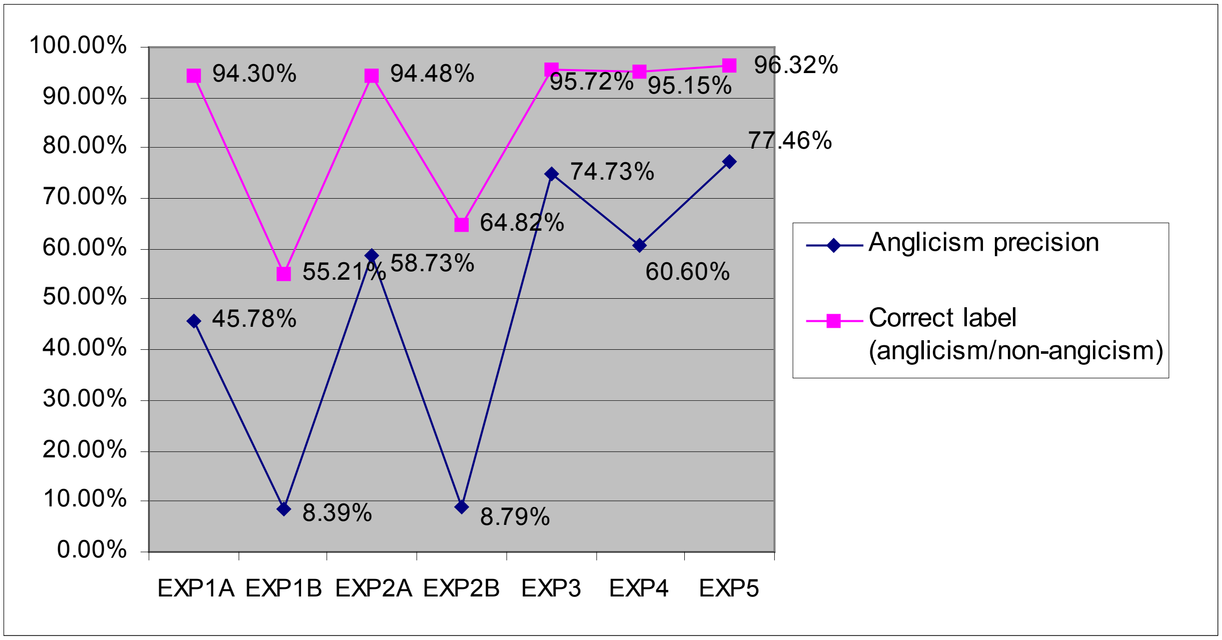

Figure 1: Overall results of experimental testing, and precision of anglicism labelling

Figure 1 shows the overall results; that is the percentage of forms that were correctly labelled as anglicism/non-anglicism by the classification tool. Experiments 1 and 2 were performed in two different stages, labelled EXP1A, EXP1B, etc. The overall performance result of the five methods varies from 55.21 per cent to 96.32 per cent correctly identified forms. The figure also shows the precision; that is the percentage of tool-classified anglicisms that were correctly labelled. This varies from 8.39 per cent to 77.46 per cent.

5.1 Experiment 1: lexicon-based method using BNC

The initial phase of this first experiment involved a lookup in a large BNC-based lexicon containing 938,972 word forms. The lexicon was the full, unlemmatised frequency list, retrieved from http://www.itri.brighton.ac.uk/~Adam.Kilgarriff/bnc-readme.html. Phase A of this experiment used exact matching, which is, of course, a non-ideal method, since it can only be expected to perform well in cases where a neologism exactly matches an English original. Since Norwegian word formation regularly involves compounding, this cannot be expected very often. Consequently, the method correctly retrieved only 38 anglicisms including *bossy* and *jackup*. Totally 563 anglicisms were identified in the manually annotated dataset, which gives a recall of 6,75 per cent. Also, these correct identifications were outnumbered by the 45 items classed as anglicisms that in fact were not, including items of non-English origin (*jujitsu*) and Norwegian-English homographs (interims). This gives a precision of 45,78 per cent, as shown in Table 1.

	Correct	Wrong	Correct %
Anglicisms	38	45	45.78%
Non-anglicisms	9392	525	94.71%

Table 1: Correctly and incorrectly tool-classified items in Experiment 1A

Note also that the classification tool undergenerates considerably, as 525 items were incorrectly classified as non-anglicisms.

In the second phase of the experiment, I modified the lexicon-based method by allowing for open-ended (fuzzy) matches, at the same time restricting the BNC reference list to words of a certain length, in order to prevent overgeneration in pattern matching against very short words. This lead to correct identification of several compounds such as *milkshakebøttene,* 'the milk shake buckets' and *croonerkul* 'crooner cool', not previously recognised. But this method also lead to overgeneration in non-anglicisms such as *skolehistoriens* 'the school history's' because, in fact, the Norwegian word for 'school', *skole*, occurs once in the BNC. Similarly, a non-anglicism like *angstdrevne* 'fear-driven' was wrongly classified because the German borrowing *angst* exists in English.

This processing proved highly time-consuming for the classification tool. I attempted to restrict the match algorithm to the available **lemmatised** BNC list of 6,318 words with more than 800 occurrences in the BNC (ftp://ftp.itri.bton.ac.uk/bnc/lemma.al.). However, despite the smaller reference lexicon, the tool actually overgenerated considerably in the identification of anglicisms, as shown in Table 2. The overall correctness percentage was no more than 55.21, as seen in Figure 1 above.

	Correct	Wrong	Correct %
Anglicisms	395	4311	8.39%
Non-anglicisms	5126	168	96.83%

Table 2: Correctly and incorrectly tool-classified items in Experiment 1B

Generally, the results of the BNC-based lexicon method were relatively poor. On the one hand, the method of exact pattern matching in the full BNC list does not capture loan words that are part of compounds. On the other hand, open-ended matching against a shorter list massively overgenerates, as seen from the 4,311 wrongly classified items in Table 2. In other words, lexicon lookup is a problematic method. However, the both versions of this experiment may be improved if the tool is enhanced with morphological decomposition of the forms to be investigated.

5.2 Experiment 2: lexicon-based method using uniquely English lexicon

For the second experiment I produced a tailored lexicon consisting of 251,063 word forms that were found to be uniquely English. That is, the tailored lexicon included those word forms in the long, unlemmatised BNC list mentioned above that were not included in a general Norwegian lexical database. This amounted to 251,000 word forms. The Norwegian lexicon used for reference is the one that was developed as part of the SCARRIE project (Scandinavian Proofreading Tools; cf. http://www.ling.uib.no/~desmedt/scarrie/). It consists of 360,000 inflectional forms.

Using the exact match method as in Experiment 1A above, the problem is again that the lexicon-based method does not capture words where the anglicism is part of a compound. However, using the tailored lexicon did reduce the amount of incorrectly classified anglicisms. Compare the anglicism figures in Table 3 with those in Table 1 above.

	Correct	Wrong	Correct %
Anglicisms	37	26	58.73%
Non-anglicisms	9411	526	94.71%

Table 3: Correctly and incorrectly tool-classified items in Experiment 2A

Similarly, if an open-ended match is used, like in Experiment 1B above, the tool incorrectly classifies many non-anglicisms as anglicisms, as seen from the figure 3,270 in Table 4. But, this is lower than the corresponding figure in Table 2 above, so the use of the tailored lexicon can be seen to have a certain effect on the overall result.

	Correct	Wrong	Correct %
Anglicisms	315	3270	8.79%
Non-anglicisms	6167	248	96.13%

Table 4: Correctly and incorrectly tool-classified items in Experiment 2B

However, on the whole, this lexicon-based method shares the problems with the method described in Experiment 1 above.

5.3 Experiment 3: chargrams method

The next experiment involved a method that was not lexicon-based but which compared each neologism with a list of so-called chargrams. The notion of ‘chargram’ denotes sequences of n characters that occur within word forms that are found in a corpus. This comparison is also based on the BNC, but in a different manner than the two previous experiments, as it compares new words with orthographical units that are statistically based letter combinations and not independent word forms. The chargram list used in the experiment was compiled from the lists produced by the Phrases in English project (http://pie.usna.edu/explorec.html). An extract of the list is shown in Figure 2.

```
ability
ably
abou
acce
acco
aced
ache
achi
achin
acin
acing
acke
acked
acki
ackin
acking
```

Figure 2: Extract of the applied chargram list

The applied chargram list contains 1074 items such as the above, consisting of 4-6 characters. The list items constitute chargrams that are **productive**, in the sense that they occur 100 times or more in the BNC, that is, in 100 or more unique word forms. Moreover, they are **uniquely English** chargrams, in the sense that they are non-existent in the Norwegian reference lexicon mentioned above.

Introducing this chargram-based method improves the overall performance of the anglicism tool considerably, yielding an overall correctness percentage of 95.72 (cf. Figure 1 above). As seen in Table 5, the tool correctly identifies 204 anglicisms, including a number of compounds containing frequent components like *action* or *country*. The recall value of this experiment was 36,23 per cent.

	Correct	Wrong	Correct %
Anglicisms	204	69	74.73%
Non-anglicisms	9368	359	96.31%

Table 5: Correctly and incorrectly tool-classified items in Experiment 3

But the tool also undergenerates by not identifying words like *soulgruppen* 'the soul group' and *surroundsound*. This happens in 69 cases, suggesting that a more comprehensive and modified list of typically English chargrams would improve the quality of this method.

5.4 Experiment 4: regular expressions

The next experiment was not based on lookup in a lexicon or a list of statistical chargrams, but the retrieval method involved manually constructed regular expressions. A set of processing rules were written with a view to identifying anglicisms, rules which were more or less intuitively constructed on the basis of general knowledge of the orthography of English. These included pattern matching against letter combinations of varying length (2-6 letters) known to occur initially, medially or finally in English words, such as the patterns *co-*, *-atio-* and *-ly*, respectively. A list of stopwords was included to prevent overgeneration.

As shown in Figure 1, the overall result of this method was similar to the results of Experiment 4, roughly 95 per cent correct labels. However, as seen from Table 6, the precision level was lower, namely 60 per cent as opposed to 74 per cent for the corresponding figure in Table 5. This gave a recall of 39 per cent.

	Correct	Wrong	Correct %
Anglicisms	223	145	60.60%
Non-anglicisms	9292	340	96.47%

Table 6: Correctly and incorrectly tool-classified items in Experiment 4

The figures for words that were incorrectly labelled as non-anglicisms were the same as in the previous experiment. This shows that manually written rules may yield results that are comparable to the chargram method.

5.5 Experiment 5: chargrams and regular expressions

The final experiment involves a combination of the hitherto most promising methods, the chargram method used in Experiment 3 and the rule-based method used in Experiment 4. The tool first identified anglicisms using the regular expressions and the stop word list identical to those described in section 5.4. Secondly, it applied pattern matching using the same list of chargrams as that described in section 5.3.

This combinatory method produced the best result so far. It gave an overall result of 95.88 per cent correctly annotated forms, which is shown in Figure 1 above. The tool classified 275 units as correctly as anglicisms, giving a precision of 75.76 per cent, as seen in Table 7.

	Correct	Wrong	Correct %
Anglicisms	275	88	75.76%
Non-anglicisms	9357	288	97.01%

Table 7: Correctly and incorrectly tool-classified items in Experiment 5

Using this method, the recall was 49 per cent.

6 Concluding remarks

In the series of experiments reported above, I have tested rule-based, lexicon-based and chargram-based methods for retrieval of anglicisms in Norwegian texts. Neither of these methods is currently able to identify all the manually identified words entirely at this stage.

Some general conclusions can be drawn from the above experiments:

- Chargram-based methods seem to outperform lexicon-based methods.
- Chargram-based methods can be improved by regular expression processing.
- Using a tailored lexicon of uniquely English word forms as reference gives a better result than using a general English lexicon.

The above experiments have shown that the optimal results are gained using combinatory methods. The most successful strategy so far is the lookup in a list of characteristically English chargrams combined with rule-based processing using regular expressions.

The described module for anglicism retrieval can still be much improved. The findings of this study have shown which methods are most successful, and which direction of future work is likely to give the best improvements of the classification tool. The manually written regular expressions can be amended using the lists of wrongly classified word forms as a basis.

In all likelihood, it will also improve the results if a finer classification of morphologically decomposed word forms is implemented. This could also enable a more precise lookup of anglicisms in a lexicon of uniquely English lemmas, where the anglicism classification applies to individual parts of words, such as compound components. Moreover, a possible future improvement is the implementation of multiword unit processing.

References

Andersen, G. (2004a) Methods and tools for automatic extraction of anglicisms in the Norwegian Newspaper Corpus. Paper presented at the 25th ICAME conference. Verona, May 2004.

Andersen, G. (2004b) Variation in the use of discourse markers in Norwegian newspaper texts. Paper presented at the 15th Sociolinguistics Symposium. Newcastle, April 2004.

Graedler, A-L & S. Johansson. (1997) *Anglisismeordboka – Engelske lånord i norsk* (Oslo: Universitetsforlaget).

Hofland, K. (2000) A Self-Expanding Corpus Based on Newspapers on the Web. *Proceedings of the Second International Language Resources and Evaluation Conference* (Paris: European Language Resources Association).

Wangensteen, B. (2002) Nettbasert nyordsinnsamling. *Språknytt*, 2/2002, 17-19.