

# COMPARING TWO LANGUAGE IDENTIFICATION SCHEMES

*Gregory Grefenstette*

Xerox Research Centre Europe  
6 chemin de Maupertuis, 38240 Meylan, France  
Gregory.Grefenstette@xrce.xerox.com

*Une des premières tapes dans le traitement automatique du langage naturel des textes reus d'une varit de sources est l'identification de la langue dans laquelle le texte a t crit. Ici, nous comparons l'utilit de deux schmas d'identification des langues. Ces techniques ne ncessitent aucune connaissance linguistique des langues traites. L'utilisation de la frquence des trigrammes semble marcher mieux que la reconnaissance de mots courts pour l'identification des langues.*

KEY WORDS *Textual Analysis, Language Identification, N-grams.*

## 1 *The language identification problem*

With the current spread of world-wide Internet access, text is available in a great number of languages other than English. Automatic treatments of these texts, for any purpose requiring natural language processing, such as WWW indexing and interrogation or providing reading aids (Bauer, Segond, Zaenen 1995), necessitate a preliminary identification of the language used. For example, morphologically-based stemming has proven important in improving information retrieval (Hull 1995) and applying language specific algorithms implies knowing the language used. Likewise, any system that involves dictionary access must identify language to perform language-specific lemmatization. This language identification problem can be seen as a specific instance of the more general problem of classification of item using attributes (Sneath and Sokal 1973).

Here, we compare two techniques for automatic language identification given machine readable text<sup>1</sup> using easily calculable attributes. One technique uses letter trigrams (sequences of three letters) and was previously described (Beesley 1988) and (Cavnar 1993) . The other techniques is based on common short words, such as those given in (Ingle 1976). Variations of these

---

<sup>1</sup>See Sibun and Spitz(1994) for a technique for identifying language of scanned documents.

---

Dan	Dut	Eng	Fre	Ger	Ita	Nor	Por	Spa	Swe
er_	en_	_th	_de	en_	_di	et_	_de	_de	en_
en_	de_	he_	es_	er_	to_	.._	de_	de_	.._
for	_de	the	de_	_de	_de	en_	os_	os_	er_
et_	et_	.,_	ent	der	di_	er_	do_	_la	et_
ing	an_	nd_	nt_	ie_	_co	_de	que	el_	tt_
_fo	n_d	ed_	_le	ich	la_	_ha	_qu	la_	_de
_af	_he	_an	e_d	sch	re_	an_	_co	que	ar_
_de	er_	and	le_	ein	ion	de_	as_	as_	.,_
nde	_va	.._	ion	che	ent	.,_	ent	ue_	fr
els	van	_to	s_d	die	e_d	det	o_	_qu	om_
lse	een	ing	e_l	ch_	le_	ar_	ue_	_co	_oc
ret	ver	to_	_la	den	o_d	_og	_a_	_en	ch_
_sa	aar	ng_	la_	nd_	ne_	og_	o_d	en_	de_
der	_ee	er_	re_	_di	no_	te_	_se	ent	och
_i_	het	_of	on_	ung	_in	han	_o_	es_	an_

Figure 1: Most frequent trigrams per language derived from the ECI Multilingual Corpus.

---

techniques are presented here. Both techniques are applied to the same test suite and their results are evaluated.

## 2 *Trigram Technique*

The trigram technique calculates the frequency of sequences of three letters in a large language sample. The idea is to capture the intuition that, for example, a word ending in *-ck* is more likely to be an English word than a French word, just as a word ending in *-ez* is more likely to be French.

For an implementation of trigram frequency, we began with the recently available ECI CD-ROM<sup>2</sup>. From this collection of texts, we collected trigram statistics, using the first million character of text in each of the following languages: Danish, Dutch, English, French, German, Italian, Norwegian, Portuguese, Spanish, and Swedish.

Each text was tokenized using the space as sole separator, and an underscore was added before and after each token in order to mark initial and terminal bigrams. All three-character sequences were counted. Figure 1 gives the most frequent trigrams derived for each language. As attributes for each language, all trigrams appearing more than 100 times were retained. Each language was characterized by between 2550 and 3560 such trigrams. The probability of a given trigram in a given language was approximated by summing the frequencies of all the trigrams retained for that language, and then dividing each frequency by this total sum.

The probabilities were used to guess the identity of a given sentence by dividing the sentence into trigrams, and calculating the probability of the sequence of trigrams for each language, assigning a minimal probability to each unseen trigram. The most probable language is chosen as the identity.

---

<sup>2</sup>See <http://www.cogsci.ed.ac.uk/elsnet/eci.html> for information on how to obtain this data.

---

Dan	Dut	Eng	Fre	Ger	Ita	Nor	Por	Spa	Swe
i	de	the	de	der	di	.	de	de	och
af	van	and	la	die	e	og	a	la	i
og	het	to	le	und	il	det	que	que	att
at	een	of		den	che	,	o	el	som
§	en	a	et	in	la	han	e	en	en
til	in	in	des	von	a	i	do	y	r
for	dat	was	les	.	in	er	da	a	p
en	is	his	du	zu	per	"	no	los	det
om	te	that	"	dem	del	p	um	del	av
der	op	I	en	,	un	til	em	se	fr
er	voor	he	un	fr		at	para	por	med
U	met	as	que	mit	non	som	com	las	den
ikke	die	had	a	das	i	var	se	con	till
eller	De	with	qui	des	si	jeg		un	har
som	zijn	it	dans	ist	le	med	os	para	de

Figure 2: Most frequent short tokens per language derived from the ECI Multilingual Corpus.

---

### 3 *Small Word Technique*

Intuitively, common words such as determiners, conjunctions and prepositions are good clues for guessing a language. These words are often short. The second language guessing method is based on these intuitions.

This method uses the same ECI corpus data as the first but derives different language attributes from it. The first million characters of text for each language was tokenized as above, and all tokens of five characters or less were extracted. These were counted for each language, and words appearing more than three times were retained. Figure 2 shows the most frequently appearing short words for each language treated. There were between 980 and 2750 such short and common words for each language. The frequencies of these words were transformed into probabilities as in the first method.

Given a new sentence to guess, the sentence is tokenized. Tokens appearing in the short word list are assigned their probabilities and tokens not in the list are assigned a minimum probability. The probability that a sentence belongs to a given language is taken as the product of the probabilities of each token.

### 4 *Test and Evaluation*

A test corpus was made for each language<sup>3</sup> by extracting the second million characters for each treated language from the ECI CD-ROM. Each test corpus was divided into sentences by

---

<sup>3</sup>There was not enough new data to test Swedish, but Swedish was still retained as a possible language while guessing sentences in other language. So each sentence presented to the guesser was picked as one out of ten languages. If all languages were equally likely, an eleventh choice of “I don’t know”, written ??? , was another alternative.

		<i>Number of Words in Sentence</i>							
		1 or 2	3 – 5	6 – 10	11 – 15	16 – 20	21 – 30	31 – 50	more than 50
		<i>Danish</i>							
trigram		92.6	97.2	97.9	99.3	99.9	99.9	99.9	99.7
short		40.5	61.6	91.8	94.8	95.5	94.3	92.5	100.0
		<i>Dutch</i>							
trigram		70.6	91.3	98.9	99.7	100.0	99.9	100.0	100.0
short		47.1	84.2	98.5	99.2	99.5	99.6	99.9	100.0
		<i>English</i>							
trigram		78.9	97.2	99.5	99.9	99.9	100.0	99.9	100.0
short		52.6	87.7	97.3	99.8	99.9	100.0	99.9	100.0
		<i>French</i>							
trigram		69.2	93.0	94.5	93.6	99.8	100.0	99.9	100.0
short		30.8	81.8	96.0	97.2	99.8	100.0	100.0	100.0
		<i>German</i>							
trigram		90.3	97.2	99.3	99.8	99.9	100.0	100.0	100.0
short		23.1	71.6	89.6	98.2	99.8	100.0	100.0	100.0
		<i>Italian</i>							
trigram		58.8	92.9	99.6	100.0	100.0	100.0	100.0	100.0
short		16.7	65.0	96.9	99.8	100.0	100.0	99.9	100.0
		<i>Norwegian</i>							
trigram		70.8	91.3	98.1	99.5	99.7	99.9	99.9	100.0
short		87.5	97.4	99.2	99.8	99.9	100.0	100.0	100.0
		<i>Portuguese</i>							
trigram		83.5	96.6	99.4	99.9	100.0	99.9	100.0	100.0
short		51.1	88.9	98.2	99.7	99.9	99.9	100.0	100.0
		<i>Spanish</i>							
trigram		73.8	86.9	97.3	99.0	99.8	99.9	99.9	100.0
short		8.1	81.5	98.8	99.7	100.0	100.0	100.0	100.0

Figure 3: Average number of Correct Language Identifications, in function of sentence length and whether trigrams or short words were used. The system had to choose among ten languages.

declaring every period followed by a word beginning with a capital letter as a full-stop.

First using the trigram attributes and then the short word attributes, each sentence was then fed into the language guesser, and we recorded what language was guessed. Figure 4 shows which languages caused confusion for which method. A breakdown of the results are shown in Figure 3. This breakdown shows that either method works well on long sentences and that trigrams are most robust for shorter sentences.

This can be expected. In shorter sentences there is a greater chance that no language-characteristic small word is used. Shorter sentences are often titles or section headings in this corpus, and these heading words often contain characteristic trigrams, even in the absence of small words which explains the better performance of the first method for sentences of 5 or fewer words. Once sentences become longer, fifteen words or more, both methods appear to work equally well.

---

<b>Trigram Method</b>				
<i>Danish</i>	<i>Dutch</i>	<i>English</i>	<i>French</i>	<i>German</i>
6657 dan	6969 dut	7856 eng	5013 fre	5902 ger
26 nor	24 ger	7 ger	28 ger	10 dut
13 ger	12 fre	3 ita	25 spa	6 por
5 ita	8 eng	2 spa	19 ita	5 fre
5 dut	4 dan	2 por	7 por	5 eng
3 swe	2 spa	2 fre	2 dut	4 dan
3 fre	2 nor	2 dut	1 nor	3 spa
3 eng	1 por		1 eng	2 swe
1 por	1 ita			2 ita
<i>Italian</i>	<i>Norwegian</i>	<i>Portuguese</i>	<i>Spanish</i>	
5880 ita	12980 nor	8699 por	5287 spa	
19 dan	128 swe	44 spa	75 por	
13 por	26 ger	19 ita	12 ita	
10 spa	23 dan	13 ger	7 ger	
9 fre	8 eng	9 fre	6 fre	
5 dut	2 ita	6 eng	2 eng	
4 ger	2 dut	2 dut	1 nor	
3 eng	1 por	1 swe	1 dut	
1 nor				

<b>Small Word Method</b>				
<i>Danish</i>	<i>Dutch</i>	<i>English</i>	<i>French</i>	<i>German</i>
6032 dan	6914 dut	7808 eng	5023 fre	5590 ger
286 nor	61 ???	39 nor	24 ita	125 ???
275 ???	17 spa	8 por	14 spa	92 fre
85 swe	9 ger	6 ger	13 ???	56 dan
25 fre	8 eng	6 ???	7 por	33 por
5 ger	5 fre	3 swe	4 swe	14 ita
3 dut	4 swe	2 fre	4 dan	11 swe
2 por	2 nor	2 dut	3 ger	7 dut
1 spa	2 dan		2 nor	5 nor
1 ita	1 por		2 dut	5 eng
1 eng				1 spa
<i>Italian</i>	<i>Norwegian</i>	<i>Portuguese</i>	<i>Spanish</i>	
5742 ita	13102 nor	8505 por	5150 spa	
97 ???	49 swe	140 ???	198 ???	
35 dan	8 dan	69 spa	23 por	
18 por	2 ita	26 ita	16 ita	
15 nor	2 ger	14 ger	3 fre	
11 ger	2 fre	12 fre	1 ger	
9 spa	2 ???	9 eng		
8 fre	1 spa	8 dut		
5 swe	1 eng	6 swe		
3 dut	1 dut	3 nor		
1 eng		2 dan		

Figure 4: Confusion matrix. Shows for each language, and each method, which languages were guessed over entire test corpus. ??? signifies that all languages were equally probable.

---

## 5 Conclusion

We have presented two techniques for identifying a language using non-linguistic corpus-derived attributes, trigrams or short words. Between 1000 and 4000 attributes were derived for each language under either method.

Both methods are easy to implement. Using short words is slightly more rapid in execution since there are less words than there are trigrams in a given sentence, and each sentence attribute contributes a multiplication to the probability calculation.

## REFERENCES

Daniel Bauer, Frederique Segond, and Annie Zaenen. 1995. Locolex: The translation rolls off your tongue. In *Proceedings of the ACH/ALLC '95*, Santa Barbara, California, July 11–15.

Kenneth R. Beesley. 1998. Language identifier: A computer program for automatic natural-language identification of on-line text. In *Language at Crossroads: Proceedings of the 29th Annual Conference of the American Translators Association*, pages 47–54. Oct 12–16.

William B. Cavnar and John M. Trenkle. 1994. N-gram-based text categorization. In *1994 Symposium On Document Analysis and Information Retrieval*, pages 161–176, University of Nevada, Las Vegas.

David A. Hull. 1996. Stemming algorithms: A case study for detailed evaluation. *JASIS*, 47(1), January. Special Issue on the Evaluation of Information Retrieval Systems.

Norman C. Ingle. 1976. A language identification table. *The Incorporated Linguist*, 15(4):98–101.

Patricia Newman. 1987. Foreign language identification: First step in the translation process. In *Proceedings of the 28th Annual Conference of the American Translators Association*, pages 159–162. October.

Penelope Sibun and A. Lawrence Spitz. 1994. Language determination: Natural language processing from scanned document images. In *ANLP'4*, pages 15–21, Stuttgart.

P. H. A. Sneath and R. R. Sokal. 1973. *Numerical Taxonomy*. W. H. Freeman, San Francisco.