Report on CLEF-2002 Experiments:
Combining Multiple Sources of Evidence

Jacques Savoy

Institut interfacultaire d'informatique, Université de Neuchâtel, Switzerland
Jacques.Savoy@unine.ch   Web site: www.unine.ch/info/clef/

Abstract. For our second participation in the CLEF retrieval tasks, our first objective was to propose better and more general stopword lists for various European languages (namely, French, Italian, German, Spanish and Finnish) along with improved, simpler and efficient stemming procedures. Our second goal was to propose a combined query-translation approach that could cross language barriers and also an effective merging strategy based on logistic regression for accessing the multilingual collection. Finally, within the Amaryllis experiment, we wanted to analyze how a specialized thesaurus might improve retrieval effectiveness.

Introduction

Based on our experiments of last year [Savoy 2002b], we participate in French, Italian, Spanish, German, Dutch and Finnish monolingual tasks in which our information retrieval approaches could work without having to rely on a dictionary. In Section 1, we improve our stopword lists and simple stemmers for the French, Italian, Spanish and German languages. For German, we also propose a new decompounding algorithm. For Dutch, we use the available stoplist and stemmer, and for the Finnish language we design a new stemmer and stopword list. In order to obtain a better overview, we evaluate our propositions using ten different retrieval schemes.

In Section 2, for the various bilingual tracks we choose to express the submitted requests in the English language, which are in turn automatically translated using five different machine translation (MT) systems and one bilingual dictionary. We study these various translations, and based on the relative merit of each translation device we investigate various combinations of them.

In Section 3, we carry out a multilingual information retrieval, investigating various merging strategies based on the results obtained during our bilingual tasks. Finally, in the last section, we present various experiments done using the Amaryllis corpus, within which a specialized thesaurus is made available in order to improve the retrieval effectiveness of the information retrieval system.

1. Monolingual indexing and search

Most European languages included in the Indo-European language family (including French, Italian, Spanish, German and Dutch) can be viewed as flectionnal languages within which polymorphs suffixes are added at the end of a flexed root. On the other hand, the Finnish language, member of the Uralic language family (together with the Turkish language), is based on a concatenative morphology in which suffixes, more or less invariable, are added to roots that are generally invariable.

Any adaptation of those indexing or search strategies available for the English language requires that general stopword lists and fast stemming procedures be developed for the other target languages. Stopword lists contain non-significant words that are removed from a document or a request before the indexing process is begun. Stemming procedures try to remove inflectional and derivational suffixes in order to conflate word variants into the same stem or root.

This first section will deal with these issues and is organized as follows: Section 1.1 contains an overview of our eight test-collections while Section 1.2 describes our general approach to building stopword lists and stemmers for use with languages other than English. In order to decompound German words, we try a simple decompounding algorithm as described in Section 1.3. Section 1.4 depicts the Okapi probabilistic model together with various vector-space models and we evaluate them using eight test-collections written in seven different languages (monolingual track).

1.1. Overview of the test-collections

The corpora used in our experiments included newspapers such as the *Los Angeles Times* (1994, English) *Le Monde* (1994, French), *La Stampa* (1994, Italian), *Der Spiegel* (1994/95, German) and *Frankfurter Rundschau* (1994, German), *NRC Handelsbald* (1994/95, Dutch), *Algemeen Dagblad* (1995/95, Dutch) and *Tidningarnas Telegrambyrå* (1994/95, Finnish). As a second source of information, we also used various articles edited by news agencies such as *EFE* (1994, Spanish), and the Swiss news agency (1994, available in French, German and Italian but without parallel translation). As shown in Table 1a and 1b, these corpora are of various sizes, with the English, German, Spanish and Dutch collections being twice the volume of the French, Italian and Finnish sources. On the other hand, the mean number of distinct indexing terms per document is relatively similar across the corpora (around 120), and this number is a little bit higher for the English collection (167.33). The Amaryllis collection contains abstracts of scientific papers written mainly in French and this corpus contains fewer distinct indexing terms per article (70.418).

	English	French	Italian	German	Spanish
Size (in MB)	425 MB	243 MB	278 MB	527 MB	509 MB
# of documents	113,005	87,191	108,578	225,371	215,738
# of distinct terms	330,753	320,526	503,550	1,507,806	528,382
Number of distinct indexing terms / document					
Mean	167.33	130.213	129.908	119.072	111.803
Standard deviation	126.315	109.151	97.602	109.727	55.397
Median	138	95	92	89	99
Maximum	1,812	1,622	1,394	2,420	642
Minimum	2	3	1	1	5
Max df	69,082	42,983	48,805	82,909	215,151
Number of indexing terms / document					
Mean	273.846	181.559	165.238	152.004	156.931
Standard deviation	246.878	164.347	130.728	155.336	82.133
Median	212	129	115	111	137
Maximum	6,087	3,923	3,763	6,407	1,003
Minimum	2	3	2	1	5
Number of queries	42	50	49	50	50
Number rel. items	821	1,383	1,072	1,938	2,854
Mean rel./request	19.548	27.66	21.878	38.76	57.08
Standard deviation	20.832	34.293	19.897	31.744	67.066
Median	11.5	13.5	16	28	27
Maximum	96 (#q:95)	177 (#q:95)	86 (#q:103)	119 (#q:103)	321 (#q:95)
Minimum	1 (#q:97,98,136)	1 (#q:121)	3 (#q:121, 132)	1 (#q:137)	3 (#q:111)

Table 1a: Test-collection statistics

When examining the number of relevant documents per request, Tables 1a and 1b show that the mean number is always greater than the median (e.g., for the English collection, there is an average of 19.548 relevant documents per query and the corresponding median is 11.5). These findings indicate that each collection contains numerous queries with a rather small number of relevant items. For each collection, we encounter 50 queries except for the Italian corpus (for which Query #120 does not have any relevant items) and the English collection (for which Query #93, #96, #101, #110, #117, #118, #127 and #132 do not have any relevant items). The Finnish corpus contains only 30 available requests while only 25 queries are included in the Amaryllis collection.

From the original documents and during the indexing process, we retained only the following logical sections in our automatic runs: <TITLE>, <HEADLINE>, <TEXT>, <LEAD>, <LEAD1>, <TX>, <LD>, <TI> and <ST>. On the other hand, we did conduct two experiments (indicated as manual runs), one with the French collection and one with the German corpus, within which we retained the following tags: for the French collection: <DE>, <KW>, <TB>, <CHA1>, <SUBJECTS>, <NAMES>, <NOM1>, <NOTE>, <GENRE>, <PEOPLE>, <SU11>, <SU21>, <GO11>, <GO12>, <GO13>, <GO14>, <GO24>, <TI01>, <TI02>, <TI03>, <TI04>, <TI05>, <TI06>, <TI07>, <TI08>, <TI09>, <ORT1>, <SOT1>, <SYE1> and <SYF1>; while for the German corpus and for one experiment, we used also the following tags: <KW> and <TB>.

From the topic descriptions we automatically removed certain phrases such as "Relevant document report …", "Find documents …", "Trouver des documents qui parlent …", "Sono valide le discussioni e le decisioni …", "Relevante Dokumente berichten …" or "Los documentos relevantes proporcionan información …".

To evaluate our approaches, we used the SMART system as a test bed for implementing the Okapi probabilistic model [Robertson 2000] as well as other vector-space models. This year our experiments were conducted on an Intel Pentium III/600 (memory: 1 GB, swap: 2 GB, disk: 6 x 35 GB).

	Dutch	Finnish	Amaryllis
Size (in MB)	540 MB	137 MB	195 MB
# of documents	190,604	55,344	148,688
# of distinct terms	883,953	1,483,354	413,262
Number of distinct indexing terms / document			
Mean	110.013	114.01	70.418
Standard deviation	107.037	91.349	31.9
Median	77	87	64
Maximum	2297	1,946	263
Minimum	1	1	5
Max df	325,188	20,803	61,544
Number of indexing terms / document			
Mean	151.22	153.73	104.617
Standard deviation	162.027	128.783	54.089
Median	101	123	91
Maximum	4510	6,117	496
Minimum	1	1	6
Number of queries	50	30	25
Number rel. items	1,862	502	2,018
Mean rel./request	37.24	16.733	80.72
Standard deviation	49.873	14.92	46.0675
Median	21	8.5	67
Maximum	301 (#q:95)	62 (#q:124)	180 (#q:25)
Minimum	4 (#q:110)	1 (#q:114)	18 (#q:23)

Table 1b: Test-collection statistics

1.2. Stopword lists and stemming procedures

In order to define general stopword lists, we used those lists already available for the English and French languages [Fox 1990], [Savoy 1999], while for the other languages we established a general stopword list by following the guidelines described in [Fox 1990]. These lists mainly contain the top 200 most frequent words included in the various collections together with articles, pronouns, prepositions, conjunctions or very frequently occurring verb forms (e.g., to be, is, has, etc.). Stopword lists used during our previous participation [Savoy 2002b] were often extended. For example for the English we used that provided by the SMART system (571 words), 431 Italian words (no change from last year), 462 French words (previously 217), 603 German words (previously 294), 351 Spanish terms (previously 272), 1,315 Dutch terms (available at CLEF Web site) and 1,134 Finnish words (these stopword lists are available at www.unine.ch/info/clef/).

After removing high frequency words, an indexing procedure uses a stemming algorithm that attempts to conflate word variants into the same stem or root. In developing this procedure for the French, Italian, German and Spanish languages, it is important to remember that these languages have more complex morphologies than does the English language [Sproat 1992]. As a first approach, our intention was to remove only inflectional suffixes such that singular and plural word forms or feminine and masculine forms conflate to the same root. More sophisticated schemes have already been proposed for the removal of derivational suffixes (e.g., "-ize", "-ably", "-ship" in the English language), such as the stemmer developed by Lovins [1968] is based on a list of over 260 suffixes, while that of Porter [1980] looks for about 60 suffixes. Figuerola [2002] for example described two different stemmers for the Spanish language, and the results show that removing only inflectional suffixes (88 different inflectional suffixes were defined) seemed to provide better retrieval levels than did removing both inflectional and derivational suffixes (this extended stemmer included 230 suffixes).

Our various stemming procedures can be found at www.unine.ch/info/clef/. This year we improved our stemming algorithms for French, within which some derivational suffixes were also removed. For the Dutch language, we use the Kraaij & Pohlmann's stemmer (ruulst.let.ruu.nl:2000/uplift/ulift.html) [Kraaij 1996]. For the Finnish language, our stemmer tries to conflate various word declinations into the same stem. Also, the Finnish language makes a distinction between partial object and whole object (e.g., "syön leilää" or "I'm eating bread" and "syön leivan" for "I'm eating the whole bread"). This aspect is not actually taken into consideration.

Finally, diacritic characters are usually not present in English collections (with some exceptions, such as "à la carte" or "résumé"); and such characters are replaced by their corresponding non-accentuated letter in the Italian, Dutch, Finnish, German and Spanish language.

1.3. Decompounding German words

Most European languages manifests other morphological characteristics that we have been considered by our approach, with compound word constructions being just one example (e.g., handgun, worldwide). In German compound words are widely used and this causes more difficulties than does English. For example, a life insurance company employee would be "Lebensversicherungsgesellschaftsangestellter" (Leben + S + versicherung + S + gesellschaft +S + angestellter for life + insurance + company + employee). Also the morphological marker ("S") is not always present (e.g., "Bankangestelltenlohn" built as Bank + angestellter + lohn (salary)). In Finnish, we also encounter similar constructions as such as "rakkauskirje" (rakkaus + kirje for love + letter) or "työviikko" (työ + viikko for work + week).

String sequence	End of previous word	Beginning of next word									
schaften	schaft	·	tion	tion	·	ern	er	·	schg	sch	g
weisen	weise	·	ling	ling	·	tät	tät	·	schl	sch	l
lischen	lisch	·	igkeit	igkeit	·	net	net	·	schh	sch	h
lingen	ling	·	lichkeit	lichkeit	·	ens	en	·	scht	sch	t
igkeiten	igkeit	·	keit	keit	·	ers	er	·	dtt	dt	t
lichkeit	lichkeit	·	erheit	erheit	·	ems	em	·	dtp	dt	p
keiten	keit	·	enheit	enheit	·	ts	t	·	dtm	dt	m
erheiten	erheit	·	heit	heit	·	ions	ion	·	dtb	dt	b
enheiten	enheit	·	lein	lein	·	isch	isch	·	dtw	dt	w
heiten	heit	·	chen	chen	·	rm	rm	·	ldan	ld	an
haften	haft	·	haft	haft	·	rw	rw	·	ldg	ld	g
halben	halb	·	halb	halb	·	nbr	n	br	ldm	ld	m
langen	lang	·	lang	lang	·	nb	n	b	ldq	ld	q
erlichen	erlich	·	erlich	erlich	·	nfl	n	fl	ldp	ld	p
enlichen	enlich	·	enlich	enlich	·	nfr	n	fr	ldv	ld	v
lichen	lich	·	lich	lich	·	nf	n	f	ldw	ld	w
baren	bar	·	bar	bar	·	nh	n	h	tst	t	t
igenden	igend	·	igend	igend	·	nk	n	k	rg	r	g
igungen	igung	·	igung	igung	·	ntr	n	tr	rk	r	k
igen	ig	·	ig	ig	·	fff	ff	f	rm	r	m
enden	end	·	end	end	·	ffs	ff		rr	r	r
isten	ist	·	ist	ist	·	fk	f	k	rs	r	s
anten	ant	·	ant	ant	·	fm	f	m	rt	r	t
ungen	ung	·	tum	tum	·	fp	f	p	rw	r	w
schaft	schaft	·	age	age	·	fv	f	v	rz	r	z
weise	weise	·	ung	ung	·	fw	f	w	fp	f	p
lisch	lisch	·	enden	end	·	schb	sch	b	fsf	f	f
ismus	ismus	·	eren	er	·	schf	sch	f	gss	g	s

Table 2: Decompounding patterns for German

According to Monz & de Rijke [2002] or [Chen 2002], including both compounds and their composite parts (only noun-noun decompositions in [Monz 2002]) in queries and documents can result in better performance while according to Molina-Salgado [2002], the decomposition of German words seems to reduce average precision.

Our approach seeks to break up those words having an initial length greater than or equal to eight characters. Moreover, decomposition cannot take place before an initial sequence [V]C, meaning that a word might begin with a series of vowels that must be followed by at least one consonant. The algorithm then seeks the occurrence of one of the models described in Table 2. For example, the last model "gss g s" indicates that when we encounter the character string "gss" the computer is allowed to cut the compound term, ending the first word with "g" and beginning the second with "s". All the models depicted in Table 2 often include letters sequences impossible to find in a simple German word such as "dtt," "fff," or "ldm". Once it has detected this pattern, the computer makes sure that the right part consists of at least four characters, potentially beginning with a series of vowels (criterion noted as [V]), followed by a CV sequence. If decomposition proves to be possible, the algorithm begins working on the right part of the decomposed word.

As an example, take the compound word "Betreuungsstelle" (meaning "care center" and made up "Betreuung" (care) and "Stelle" (center, place)). This word is definitely more than seven characters long. Once this has been verified, the computer begins searching for substitution models for the third character. The computer will find a match with the last model described in Table 2, and form the words "Betreuung" and "Stelle." This break is validated because the second word has a length greater than four characters. This term also meets criterion [V]CV and finally, given that the term "Stelle" has less than eight letters, the computer will not attempt to continue decomposing this term.

1.4. Indexing and searching strategy

In order to obtain a broader view of the relative merit of various retrieval models, we first adopted a binary indexing scheme within which each document (or request) is represented by a set of keywords, without any weight. To measure the similarity between documents and requests, we count the number of common terms, computed according to the inner product (retrieval model denoted "doc=bnn, query=bnn" or "bnn-bnn"). For document and query indexing however binary logical restrictions however are often too limiting. In order to weight the presence of each indexing term in a document surrogate (or in a query), we may take account of the term occurrence frequency which allows for better term distinction and increases indexing flexibility (retrieval model notation: "doc=nnn, query=nnn" or "nnn-nnn").

bnn	$w_{ij} = 1$	nnn	$w_{ij} = tf_{ij}$
ltn	$w_{ij} = (\ln(tf_{ij}) + 1) \cdot idf_j$	atn	$w_{ij} = idf_j \cdot [0.5 + 0.5 \cdot tf_{ij} / \max tf_{i.}]$
nfn	$w_{ij} = \ln\left[n / df_j \right]$	npn	$w_{ij} = tf_{ij} \; \ln\left[\left(n - df_j\right) / df_j \right]$
Okapi	$w_{ij} = \left(\left(k_1 + 1\right) \; tf_{ij}\right) / \left(K + tf_{ij}\right)$	Lnu	$w_{ij} = \dfrac{\left(\left(1 + \ln(tf_{ij})\right) / \left(1 + pivot\right) \right)}{(1 - slope) \; pivot + slope \; nt_i}$
lnc	$w_{ij} = \dfrac{\ln(tf_{ij}) + 1}{\sqrt{\sum_{k=1}^{t} \left(\ln(tf_{ik}) + 1\right)^2}}$	ntc	$w_{ij} = \dfrac{tf_{ij} \; idf_j}{\sqrt{\sum_{k=1}^{t} \left(tf_{ik} \; idf_k\right)^2}}$
dtc	$w_{ij} = \left(\ln\left(\ln(tf_{ij}) + 1\right) + 1\right) \; idf_j \Big/ \sqrt{\sum_{k=1}^{t} \left(\left(\ln\left(\ln(tf_{ik}) + 1\right) + 1\right) \; idf_k\right)^2}$		
ltc	$w_{ij} = \left(\ln(tf_{ij}) + 1\right) \; idf_j \Big/ \sqrt{\sum_{k=1}^{t} \left(\left(\ln(tf_{ik}) + 1\right) \; idf_k\right)^2}$		
dtu	$w_{ij} = \dfrac{\left(\left(1 + \ln\left(1 + \ln(tf_{ij})\right)\right) \; idf_j \Big/ 1 + pivot \right)}{(1 - slope) \; pivot + slope \; nt_i}$		

Table 3: Weighting schemes

Those terms however that do occur very frequently in the collection are not considered very helpful in distinguishing between relevant and non-relevant items. Thus we might count their frequency in the collection, or more precisely the inverse document frequency (denoted by idf), resulting in more weight for sparse words and less weight for more frequent ones. Moreover, a cosine normalization could prove beneficial and each indexing weight could vary within the range of 0 to 1 (retrieval model notation: "ntc-ntc", Table 3 depicts the exact weighting formulation).

Other variants may also be created, especially if we consider the occurrence of a given term in a document is a rare event. Thus, it may be a good practice to give more importance to the first occurrence of this word as compared to any successive or repeating occurrences. Therefore, the tf component may be computed as 0.5 + 0.5 · [tf / max tf in a document] (retrieval model denoted "doc=atn").

Finally, we should consider that a term's presence in a shorter document provides stronger evidence than it does in a longer document. To account for this, we integrate document length within the weighting formula, leading to more complex IR models; for example, the IR model denoted by "doc=Lnu" [Buckley 1996], "doc=dtu" [Singhal 1999]. Finally for CLEF-2002, we also conducted various experiments using the Okapi probabilistic model [Robertson 2000] within with $K = k_1 \cdot [(1 - b) + b \cdot (l_i / avdl)]$, representing the ratio between the length of D_i measured by l_i (sum of tf_{ij}) and the collection mean noted by advl.

In our experiments, the constants b, k_1, advl, pivot and slope are fixed according to values listed in Table 4. To evaluate the retrieval performance of these various IR models, we adopted the non-interpolated average precision (computed on the basis of 1,000 retrieved items per request by the TREC-EVAL program), allowing for both precision and recall using a single number.

Language	b	k_1	advl	pivot	slope
English	0.8	2	900	100	0.1
French	0.7	2	750	100	0.1
Italian	0.6	1.5	800	100	0.1
Spanish	0.5	1.2	300	100	0.1
German	0.55	1.5	600	125	0.1
Dutch	0.9	3.0	600	125	0.1
Finnish	0.75	1.2	900	125	0.1
Amaryllis	0.7	2	160	30	0.2

Table 4: Parameter setting for the various test-collections

Given that French, Italian and Spanish morphology is comparable to that of English, we decided to index French, Italian and Spanish documents based on word stems. For the German, Dutch and Finnish languages and their more complex compounding morphology, we decided to use a 5-gram approach [McNamee 2002]. However, contrary to [McNamee 2002], our generation of 5-gram indexing terms does not span word boundaries. This value of 5 was chosen because it performed better with the CLEF-2000 corpora [Savoy 2001a]. Using this indexing scheme, the compound «das Hausdach» (the roof of the house) will generate the following indexing terms: «das», «hausd», «ausda», «usdac» and «sdach».

Our evaluation results as reported in Tables 5 show that the Okapi probabilistic model performs best with the use of five different languages. In the second position, we usually find the vector-space model "doc=Lnu, query=ltc" and in the third "doc=dtu, query=dtc". Finally, the traditional tf-idf weighting scheme ("doc=ntc, query=ntc") does not exhibit very satisfactory results, and the simple term-frequency weighting scheme ("doc=nnn, query=nnn") or the simple coordinate match ("doc=bnn, query=bnn") results in poor retrieval performance.

	Average precision			
Query T-D Model	English 42 queries	French 50 queries	Italian 49 queries	Spanish 50 queries
doc=Okapi, query=npn	**50.08**	**48.41**	**41.05**	**51.71**
doc=Lnu, query=ltc	48.91	46.97	39.93	49.27
doc=dtu, query=dtc	43.03	45.38	39.53	47.29
doc=atn, query=ntc	42.50	42.42	39.08	46.01
doc=ltn, query=ntc	39.69	44.19	37.03	46.90
doc=ntc, query=ntc	27.47	31.41	29.32	33.05
doc=ltc, query=ltc	28.43	32.94	31.78	36.61
doc=lnc, query=ltc	29.89	33.49	32.79	38.78
doc=bnn, query=bnn	19.61	18.59	18.53	25.12
doc=nnn, query=nnn	9.59	14.97	15.63	22.22

Table 5a: Average precision of various indexing and searching strategies (monolingual)

For the German language, we determined that 5-gram indexing, decompounded indexing and word-based document representation methods to be distinct and independent sources of evidence for German language document content. We therefore decided to combine these three indexing schemes and to do so we normalized similarity values obtained by each document extracted from these three separate retrieval models, according to Equation 1 (see

Section 3). The resulting average precision for these four approaches is shown in Table 5b, thus demonstrating how the combined model usually results in better retrieval performance.

	Average precision			
Query T-D Model	German words 50 queries	German decompounded 50 queries	German 5-gram 50 queries	German combined (Eq. 1) 50 queries
doc=Okapi, query=npn	**37.39**	**37.75**	**39.83**	**41.25**
doc=Lnu, query=ltc	36.41	36.77	36.91	39.79
doc=dtu, query=dtc	35.55	35.08	36.03	38.21
doc=atn, query=ntc	34.48	33.46	37.90	37.93
doc=ltn, query=ntc	34.68	33.67	34.79	36.37
doc=ntc, query=ntc	29.57	31.16	32.52	32.88
doc=ltc, query=ltc	28.69	29.26	30.05	31.08
doc=lnc, query=ltc	29.33	29.14	29.95	31.24
doc=bnn, query=bnn	17.65	16.88	16.91	21.30
doc=nnn, query=nnn	14.87	12.52	8.94	13.49

Table 5b: Average precision of various indexing and searching strategies (German collection)

It was observed that pseudo-relevance feedback (blind-query expansion) seems to be a useful technique for enhancing retrieval effectiveness. In this study, we adopted Rocchio's approach [Buckley 1996] with $\alpha = 0.75$, $\beta = 0.75$ whereby the system was allowed to add m terms extracted from the n best ranked documents from the original query. To evaluate this proposition, we used the Okapi probabilistic model and we enlarged the query by 10 to 20 terms provided by the 5 or 10 best-retrieved articles. The results depicted in Table 6a and 6b indicate that the optimal parameter setting seems to be collection-dependant. Moreover, performance improvement seems also to be collection dependant (or language dependant) with no improvement for the English corpus yet an increase of 8.55% for the Spanish corpus (from an average precision of 51.71 to 56.13), 9.85% for the French corpus (from 48.41 to 53.18), 12.91% for the Italian language (41.05 to 46.35) and 13.26% for the German collection (from 41.25 to 46.72, combined model, Table 6b).

	Average precision			
Query T-D Model	English 42 queries	French 50 queries	Italian 49 queries	Spanish 50 queries
doc=Okapi, query=npn	**50.08**	48.41	41.05	51.71
5 docs / 10 best terms	49.54	53.10	45.14	55.16
5 docs / 15 best terms	48.68	**53.18**	46.07	54.95
5 docs / 20 best terms	48.62	53.13	**46.35**	54.41
10 docs / 10 best terms	47.77	52.03	45.37	55.94
10 docs / 15 best terms	46.92	52.75	46.18	56.00
10 docs / 20 best terms	47.42	52.78	45.87	**56.13**

Table 6a: Average precision using blind-query expansion

	Average precision			
Query T-D Model	German words 50 queries	German decompounded 50 queries	German 5-gram 50 queries	German combined (Eq. 1) 50 queries
doc=Okapi, query=npn	37.39	37.75	39.83	41.25
# docs / # terms	5 / 40 42.90	5 / 40 42.19	10 / 200 45.45	**46.72**
# docs / # terms	5 / 40 **42.90**	5 / 40 **42.19**	5 / 300 **45.82**	46.27

Table 6b: Average precision using blind-query expansion

This year, we also participated in the Dutch and Finnish monolingual tasks, the results of which are depicted in Table 7, and the average precision of the Okapi model using blind-query expansion is given in Table 8. For these two languages, we also applied or combined an indexing model based on 5-gram indexing and word-based document representations. While for the Dutch language, our combined model seems to enhance the retrieval effectiveness, for the Finnish language it does not. This however was a first trial for our proposed stemmer and it seemed to improve the average precision over a baseline trial without stemming procedure (Okapi model, unstemmed 23.04, with stemming 30.45, an improvement of +32.16%).

	Average precision					
Query T-D	Dutch	Dutch	Dutch	Finnish	Finnish	Finnish
	word	5-gram	combined	word	5-gram	combined
Model	50 queries	50 queries	50 queries	30 queries	30 queries	30 queries
doc=Okapi, query=npn	**42.37**	**41.75**	**44.56**	30.45	**38.25**	**37.51**
doc=Lnu, query=ltc	42.57	40.73	44.50	27.58	36.07	36.83
doc=dtu, query=dtc	41.26	40.59	43.00	**30.70**	36.79	36.47
doc=atn, query=ntc	40.29	40.34	41.89	29.22	37.26	36.51
doc=ltn, query=ntc	38.33	38.72	40.24	29.14	35.28	35.31
doc=ntc, query=ntc	33.35	34.94	36.41	25.21	30.68	31.93
doc=ltc, query=ltc	32.81	31.24	34.46	26.53	30.85	33.47
doc=lnc, query=ltc	31.91	29.67	34.18	24.86	30.43	31.39
doc=bnn, query=bnn	18.91	20.87	23.52	12.46	14.55	18.64
doc=nnn, query=nnn	13.75	10.48	12.86	11.43	14.69	15.56

Table 7: Average precision of various indexing and searching strategies (Dutch and Finnish corpora)

	Average precision					
Query T-D	Dutch	Dutch	Dutch	Finnish	Finnish	Finnish
	word	5-gram	combined	word	5-gram	combined
Model	50 queries	50 queries	50 queries	30 queries	30 queries	30 queries
doc=Okapi, query=npn	42.37	41.75	44.56	30.45	38.25	37.51
# docs / # terms	5/60 47.86	5/75 45.09	48.78	5/60 31.89	5/75 40.90	39.33
# docs / # terms	5/100 **48.84**	10/150 **46.29**	**49.28**	5/15 **32.36**	5/175 **41.67**	**40.11**

Table 8: Average precision using blind-query expansion

In the monolingual track, we submitted six runs along with their corresponding descriptions, as listed in Table 9. Four of them were fully automatic using the request's Title and Descriptive logical sections, while the last three used more other document sections, based on the request's Title, Descriptive and Narrative sections. In these last three runs, two were labeled "manual" because we used logical sections containing manually assigned index terms. For all other runs however we did not use any manual intervention during the indexing and retrieval procedures.

Run name	Language	Query	Form	Model	Query expansion	average
UniNEfr	French	T-D	automatic	Okapi	no expansion	48.41
UniNEit	Italian	T-D	automatic	Okapi	10 best docs / 15 terms	46.18
UniNEes	Spanish	T-D	automatic	Okapi	5 best docs / 20 terms	54.41
UniNEde	German	T-D	automatic	combined	5/40 word, 10/200 5-gra.	46.72
UniNEnl	Dutch	T-D	automatic	combined	5/60 word, 5/75 5-gram	48.78
UniNEfi1	Finnish	T-D	automatic	Okapi	5 best docs / 75 terms	40.90
UniNEfi2	Finnish	T-D	automatic	combined	5/60 word, 5/75 5-gram	39.33
UniNEfrtdn	French	T-D-N	manual	Okapi	5 best docs / 10 terms	59.19
UniNEestdn	Spanish	T-D-N	automatic	Okapi	5 best docs / 40 terms	60.51
UniNEdetdn	German	T-D-N	manual	combined	5/50 word, 10/300 5-gram	49.11

Table 9: Official monolingual run descriptions

2. Bilingual information retrieval

In order to overcome language barriers, we based our approach on free and readily available translation resources that automatically translate queries into the desired target language. More precisely, the original queries were written in English and we used no parallel or aligned corpora to derive statistically or semantically related words in the target language. Section 2.1 describes our combined strategy for cross-lingual retrieval while Section 2.2 provides some examples of translation errors.

This year, we used five machine translation systems, namely SYSTRAN™ (babel.altavista.com/translate.dyn), GOOGLE.COM (www.google.com/language_tools), FREETRANSLATION.COM (www.freetranslation.com), INTERTRAN (www.tranexp.com:2000/InterTran) and REVERSO ONLINE (translation2.paralink.com). As bilingual dictionary we used the BABYLON system (www.babylon.com).

2.1. Query automatic translation

In order to develop a fully automatically approach, we chose to translate the requests using five different machine translation (MT) systems. We also translated query terms word-by-word using the BABYLON bilingual dictionary, provides not only one but several terms for the translation for each word submitted. In our experiments, we decided to pick the first translation available (labeled "baby1"), the first two terms (labeled "baby2") or the first three available translations (labeled "baby3").

	Average precision					
Query T-D \ Language Translation tools	French	Italian	Spanish	German word	German decomp.	German 5-gram
Original queries	48.41	41.05	51.71	37.39	37.75	39.83
Systran	42.70	32.30	38.49	28.75	28.66	27.74
Google	42.70	32.30	38.35	28.07	26.05	27.19
FreeTranslation	40.58	**32.71**	40.55	28.85	**31.42**	27.47
InterTran	33.89	30.28	37.36	21.32	21.61	19.21
Reverso	39.02	N/A	**43.28**	**30.71**	30.33	**28.71**
Babylon 1	**43.24**	27.65	39.62	26.17	27.66	28.10
Babylon 2	37.58	23.92	34.82	26.78	27.74	25.41
Babylon 3	35.69	21.65	32.89	25.34	26.03	23.66
Comb 1	46.77	33.31	44.57	34.32	34.66	32.75
Comb 2	48.02	34.70	**45.63**	35.26	34.92	32.95
Comb 2b	48.02		45.53	35.09	34.51	32.76
Comb 3	**48.56**	34.98	45.34	34.43	34.37	**33.34**
Comb 3b	48.49	35.02	45.34	34.58	34.43	32.76
Comb 3b2				**35.41**	**35.13**	33.25
MT 2		**35.82**				
MT 3	44.54	35.57	44.32	33.53	33.05	31.96
All	47.94	35.29	44.25	34.52	34.31	32.79
MT all	46.83	35.68	44.25	33.80	33.51	31.66
Comb 1	Rever-baby1	Free-baby1	Rever-baby1	Reverso-baby1		
Comb 2	Reverso systran-baby1	Free-google baby1	Rever-systran baby1	Reverso-systran-baby1		
Comb 2b	Reverso google-baby1		Rever-google baby1	Reverso-google-baby1		
Comb 3	Reverso-free google-baby1	Free-google inter-baby1	Free-google rever-baby1	Reverso-systran-inter-baby1		
Comb 3b	Reverso-inter google-baby1	Free-google systran-baby1	Free-google rever-baby2	Reverso-google-inter-baby1		
Comb 3b2				Reverso-google-inter-baby2		
MT 2		Free-google				
MT 3	Reverso systran-google	Free-google inter	Free-google reverso	Reverso-inter-systran		

Table 10: Average precision of various query translation strategies (Okapi model)

The first part of Table 10 lists the average precision for each translation devices used along the performance achieved by manually translated requests. For German, we also reported the retrieval effectiveness achieved by the three difference approach, namely using words as indexing terms, decompounding the German words according to our approach and the 5-grams model. While the REVERSO system seems to be the better choice for German and Spanish, FREETRANSLATION is the best choice for Italian and BABYLON 1 the best for French.

In order to improve search performance, we tried combining different machine translation systems with the bilingual dictionary approach. In this case, we formed the translated query by concatenating the different translations provided by the various approaches. Thus the column header "Comb 1", we combined one machine translation system with the bilingual dictionary ("baby1"). Similarly, under columns "Comb 2" or "Comb 2b," we listed the results of two machine translation approaches and three machine translation systems under column headings "Comb 3", "Comb 3b" or "Comb 3b2". With the exception of the performance under "Comb 3b2," we also included terms provided by the "baby1" dictionary look-up in the translated requests. In columns "MT 2" and "MT 3," we evaluated the combination of two and three machine translation systems respectively. Finally, we could also combine all translation sources (under heading "All") or all machine translation approaches under the heading "MT all."

Since the performance of each translation device depends on the target language, in the lower part of Table 10 we included the exact specification for each of the combined runs. For the German language, for each of the three indexing models, we used the same combination of translation resources. From an examination of the retrieval effectiveness of our various combined approaches listed in the middle part of Table 10, a clear recommendation cannot be made. Overall, it seems better to combine two or three machine translation systems with the bilingual dictionary approach ("baby1"). However, combining the five machine translation systems (heading "MT all") or all translation tools (heading "All") does not result in a very effective performance.

	Average precision				
Query T-D	French	French	French	Italian	Italian
	UniNEfrBi	UniNEfrBi2	UniNEfrBi3	UniNEitBi	UniNEitBi2
Combined	Comb 3b	MTall+baby2	MT all	Comb 2	Comb 3
Expand # docs / # terms	5 / 20	5 / 40	10 / 15	10 / 60	10 / 100
Corrected	**51.64**	50.79	48.49	38.50	**38.62**
Official	49.35	48.47	46.20	37.36	37.56
Query T-D	Spanish	Spanish	Spanish	German	German
	UniNEesBi	UniNEesBi2	UniNEesBi3	UniNEdeBi	UniNEdeBi2
Combined	MT 3	Comb 3b	Comb 2	Comb 3b2 & Comb 3	
Expand # docs / # terms	10 / 75	10 / 100	10 / 75	5 / 100 & 5 / 300	
Corrected	50.67	**50.95**	50.93	**42.89**	42.11
Official	47.63	47.86	47.84	41.29	40.42

Table 11: Average precision and description of our official bilingual runs (Okapi model)

Table 11 lists the exact specifications of our various bilingual runs. However, when submitting our official results, we used the wrong numbers for Query # 130 and # 131 (we switched these two query numbers). Thus, both requests have an average precision 0.00 in our official results and we reported the corrected performance in Tables 11 and 13 (multilingual runs).

2.2. Examples of failures

In order to obtain a preliminary picture of the automatic translation approach's underlying difficulties, we analyzed some queries through comparing translations produced by our six machine-based tools with the request formulation written by a human being (examples are given in Table 12). As a first example, the title of Query #113 is "European Cup". In this case, the term "cup" was analyzed as a teacup by all automatic translation tools, resulting in the French translations "tasse" or "verre" (or "tazza" in Italian, "Schale" in German ("Pokal" can be viewed as a correct translation alternative) and "taza" or "Jícara" (small teacup) in Spanish).

In Query #118 ("Finland's first EU Commissioner"), the machine translation systems failed to give the appropriate Spanish term "comisario" for "Commissioner" but returned "comisión" (commission) or "Comisionado" (adjective relative to commission). For this same request number, the manually translated query seemed to contain a spelling error in Italian ("commi**s**ario" instead of "commi**ss**ario"). For the same request, the translation given in German "Beauftragter" (delegate) does not correspond to the appropriate term "Kommissar" (more the missing "-" in the translation "EUBEAUFTRAGTER").

Other examples: for Query #94 ("Return of Solzhenitsyn") which is translated manually in German ("Rückkehr Solschenizyns"), our automatic translation systems fail to translate the proper noun (returning "Solzhenitsyn" instead of "Solschenizyns"). Query #109 ("Computer Security") is translated manually Spanish as "Seguridad Informática" and our various translations devices return different terms for "Computer" (e.g., "Computadora", "Computador", or "ordenador") but not the word "Informática".

<num> C113 (query translations failed in French, Italian, German and Spanish)
<EN-title> European Cup
<FR-title manually translated>   Coupe d'Europe de football
<FR-title FREETRANSLATION> Tasse européenne
<FR-title BYBYLON 1> Européen verre
<FR-title BYBYLON 2> Européen résident de verre tasse
<FR-title BYBYLON 3> Européen résident de l'Europe verre tasse coupe

<IT-title manually translated>   Campionati europei
<IT-title SYSTRAN> Tazza Europea
<IT-title GOOGLE> Tazza Europea

<GE-title manually translated>   Fussballeuropameisterschaft
<GE-title SYSTRAN> Europäische Schale
<GE-title REVERSO> Europäischer Pokal

<ES-title manually translated>   Eurocopa
<ES-title INTERTRAN> Europea Jícara
<ES-title REVERSO> Taza europea

<num> C118 (query translations failed in Italian, German and Spanish)
<EN-title> Finland's first EU Commissioner.
<IT-title manually translated>   Primo commisario europeo per la Finlandia
<IT-title GOOGLE> Primo commissario dell'Eu della Finlandia.
<IT-title FREETRANSLATION> Finlandia primo Commissario di EU.

<GE-title manually translated>   Erster EU-Kommissar aus Finnland
<GE-title GOOGLE> Finnlands erster EUBEAUFTRAGTER.
<GE-title REVERSO> Finlands erster EG-Beauftragter

<ES-title manually translated>   Primer comisario finlandés de la UE
<ES-title GOOGLE> Primera comisión del EU de Finlandia.
<ES-title REVERSO> El primer Comisionado de Unión Europea de Finlandia.

Table 12: Examples of unsuccessful query translations

3. Multilingual information retrieval

Using our combined approach to automatically translate a query, we were able to search a document collection for a request written in English. This stage however represents only the first step in a proposal for multi-language information retrieval systems. We also need to investigate situations where users write a request in English in order to retrieve pertinent documents in English, French, Italian, German and Spanish. To deal with this multi-language barrier, we divided our document sources according to language and thus formed five different collections. After searching in these corpora and obtaining five results lists, we needed to merge them in order to provide users with a single list of retrieved articles.

Recent works have suggested various solutions to merging the separate result list obtained from different collections or distributed information services. As a first approach, we will assume that each collection contains approximately the same number of pertinent items and that the distribution of the relevant documents is similar across the result lists. Based solely on the rank of the retrieved records, we can interleave the results in a round-robin fashion. According to previous studies [Voorhees 1995], the retrieval effectiveness of such an interleaving scheme is around 40% below that achieved from a single retrieval scheme working with a single huge collection, representing the entire set of documents.

To take account of the document score computed for each retrieved item (or the similarity value between the retrieved record and the request, denoted score rsv_j), we might formulate the hypothesis that each collection is searched by the same or a very similar search engine and that the similarity values are therefore directly comparable [Kwok 1995]. Such a strategy, called raw-score merging, produces a final list sorted by the document score computed by each collection. However, collection-dependent statistics in document or query weights may vary widely among collections, and therefore this phenomenon may invalidate the raw-score merging hypothesis.

To account for this fact, we might normalize the document scores within each collection by dividing them by the maximum score (i.e. the document score of the retrieved record in the first position). As a variant of this normalized score merging scheme, Powell *et al.* [2000] suggest normalizing the document score rsv_j according to the following formula:

$$rsv'_j = \left(rsv_j - rsv_{min}\right) / \left(rsv_{max} - rsv_{min}\right) \qquad (1)$$

in which rsv_j is the original retrieval status value (or document score), and rsv_{max} and rsv_{min} are the maximum and minimum document score values that a collection could achieve for the current request. In this study, the rsv_{max} is given by the document score achieved by the first retrieved item and the retrieval status value obtained by the 1000th retrieved record gives the value of rsv_{min}.

As a fourth strategy, we might use the logistic regression [Flury 1997, Chapter 7] to predict the probability of a binary outcome variable, according to a set of explanatory variables. Based on this statistical approach, Le Calvé and Savoy [2000] and Savoy [2002a] described how to predict the probability of relevance of those documents retrieved by different retrieval schemes or collections. The resulting estimated probabilities would be predicted according to both the original document score rsv_i and the logarithm of the $rank_i$ attributed to the corresponding document D_i. Based on these estimated relevance probabilities, we sorted the records retrieved from separate collections in order to obtain a single ranked list. However, in order to estimate the underlying parameters, this approach requires a training set, in this case the CLEF-2001 topics and their relevance assessments.

$$\text{Prob}\left[D_i \text{ is rel} \mid rank_i, rsv_i\right] = \frac{e^{\alpha+\beta_1 \ln(rank_i)+\beta_2\ rsv_i}}{1 + e^{\alpha+\beta_1 \ln(rank_i)+\beta_2\ rsv_i}} \qquad (2)$$

within which $rank_i$ denotes the rank of the retrieved document D_i, ln() is the natural logarithm, and rsv_i is the retrieval status value (or document score) of the document D_i. In this equation, the coefficients α, β_1 and β_2 are unknown parameters that are estimated according the method of the maximum likelihood (the required computations have been done with the S language).

	Average precision				
Query T-D	English	French	Italian	Spanish	German
	42 queries	50 queries	49 queries	50 queries	50 queries
		UniNEfrBi	UniNEitBi	UniNEesBi	UniNEdeBi
	50.08	51.64	38.50	50.67	42.89
Multilingual	Round-robin	Raw-score	Eq. 1	Log ln($rank_i$)	Log reg Eq.2
50 queries	34.27	33.83	36.62	36.10	**39.49**
	English	French	Italian	Spanish	German
	42 queries	50 queries	49 queries	50 queries	50 queries
		UniNEfrBi2	UniNEitBi2	UniNEesBi2	UniNEdeBi2
	50.08	50.79	38.62	50.95	42.11
Multilingual	Round-robin	Raw-score	Eq. 1	Log ln($rank_i$)	Log reg Eq.2
50 queries	33.97	33.99	36.90	35.59	39.25

Table 13: Average precision using various merging strategies based on automatically translated queries

When searching in multi-lingual corpora using Okapi, the round-robin scheme or the raw-score merging strategy provide very similar retrieval performances (see Table 13). The normalized score merging based on Equation 1 shows an enhancement over the round-robin approach (36.62 vs. 34.27, an improvement of +6.86% in our first experiment, and 36.90 vs. 33.97, +8.63% in our second run). Using our logistic model with only the rank as explanatory variable (or more precisely the ln($rank_i$), performance depicted under the label "Log ln($rank_i$)"), the resulting average precision is lower than the normalized score merging. When merging the result lists based on the logistic regression approach (using both the rank and the document score as explanatory variables) presents the best average precision.

Query T-D	UniNEm1	UniNEm2	UniNEm3	UniNEm4	UniNEm5
	Equation 1	Log reg Eq.2	Equation 1	Log reg Eq.2	Equation 1
Corrected	36.62	**39.49**	36.90	39.25	35.97
Official	34.88	37.83	35.12	37.56	35.52

Table 14: Average precision obtained with our official multilingual runs

Our official and corrected results are shown in Table 14 while some statistics about the number of documents provided by each collection are given in Table 15. From this data, we can see that the normalized score merging (UniNEm1) extracts more documents for the English corpus (in mean 24.94 items) than the logistic regression model (UniNEm2 where in mean 11.44 documents are coming from the English collection). Moreover, the logistic regression scheme takes more documents from the Spanish and German collections Finally, we can see that the percentage of relevant items is relatively similar when comparing CLEF01 and CLEF02 test-collections.

Statistics \ Language	English	French	Italian	Spanish	German
UniNEm1, based on the top 100 retrieved documents for each query					
Mean	24.94	16.68	19.12	23.8	15.46
Median	23.5	15	18	22	15
Maximum	60 (q#:101)	54 (q#:110)	45 (q#:136)	70 (q#:121)	54 (q#:116)
Minimum	4 (q#:108)	5 (q#:97,123)	5 (q#:93,114)	6 (q#:98,110)	2 (q#:139)
Standard deviation	13.14	9.26	9.17	14.15	9.79
UniNEm2, based on the top 100 retrieved documents for each query					
Mean	11.44	15.58	16.18	34.3	22.5
Median	9	14	16	34.5	19
Maximum	33 (q#:92)	38 (q#:110)	28 (q#:108)	62 (q#:91)	59 (q#:116)
Minimum	1 (q#:135)	6 (q:102,123)	8 (q#:114)	10 (q#:116)	4 (q#:91)
Standard deviation	6.71	7.49	5.18	10.90	11.90
% relevant items CLEF02	10.18%	17.14%	13.29%	35.37%	24.02%
% relevant items CLEF01	10.52%	14.89%	15.31%	33.10%	26.17%

Table 15: Statistics about the merging schemes based on the top 100 retrieved documents for each query

4. Amaryllis experiments

For the Amaryllis experiments, we wanted to determine whether a specialized thesaurus might improve the retrieval effectiveness over a baseline, ignoring term relationships. From the original documents and during the indexing process, we retained only the following logical sections in our runs: <text>, <ti>, <ab>, <mc>, <kw>.

<RECORD>
<TERMFR> Analyse de poste
<TRADENG> Station Analysis
…
<RECORD>
<TERMFR> Bureau poste
<TRADENG> Post offices
<RECORD>
<TERMFR> Bureau poste
<TRADENG> Post office
…
<RECORD>
<TERMFR> Isolation poste électrique
<TRADENG> Substation insulation
…
<RECORD>
<TERMFR> Caserne pompier
<TRADENG> Fire houses
<SYNOFRE1> Poste incendie
…
<RECORD>
<TERMFR> Habitacle aéronef
<TRADENG> Cockpits (aircraft)
<SYNOFRE1> Poste pilotage
…

<RECORD>
<TERMFR> La Poste
<TRADENG> Postal services
…
<RECORD>
<TERMFR> Poste conduite
<TRADENG> Operation platform
<SYNOFRE1> Cabine conduite
…
<RECORD>
<TERMFR> POSTE DE TRAVAIL
<TRADENG> WORK STATION
<RECORD>
<TERMFR> Poste de travail
<TRADENG> Work Station
<RECORD>
<TERMFR> Poste de travail
<TRADENG> Work station
<RECORD>
<TERMFR> Poste de travail
<TRADENG> workstations
<SYNOFRE1> Poste travail
…

Table 16: Sample of various entries under the word "poste" in the Amaryllis thesaurus

From the given thesaurus, we have extracted 126,902 terms having a relationship with one or more terms (the thesaurus owns 173,946 entries delimited by the tags <RECORD> … </RECORD>, however only 149,207 entries have at least one relationship with another term. From these 149,207 entries, we found 22,305 multiple entries (that are removed, as for example, the term "Poste de travail" or "Bureau poste" in Table 16). In building our thesaurus, we removed the accents, wrote all terms in lowercase, and ignored numbers and terms given between parenthesis. For example, the word "poste" appears in 49 records (usually as part of a compound entry in the <TERMFR> field).

From our 126,902 entries, we counted 107,038 TRADEENG relationships, 14,590 SYNOFRE1, 26,772 AUTOP1 relationships and 1,071 VAUSSI1 relationships (see examples given in Table 16). In a first set of experiments, we did not use this thesaurus and we used the Title and Descriptive logical sections of the requests (second column of Table 17a) or the Title, Descriptive and Narrative parts of the queries (last column of Table 17a). In a second set of experiments, we included all related words that could be found in the thesaurus using only the search keywords (average precision depicted under the label "Qthes"). In a third experiment, we enlarged only document representatives using our thesaurus (performance shown under column heading "Dthes"). In a last experiment, we take account for related words found in the thesaurus only for document surrogates and under the additional condition that such relationship can be found with at least three terms (e.g. "moteur à combustion" is a valid candidate but not single term like "moteur"). On the other hand, we also included in the query all relationships that can be found using the search keywords (performance shown under the column heading "Dthes3Qthes").

	Average precision				
Query	Amaryllis T-D	Amaryllis T-D Qthes	Amaryllis T-D Dthes	Amaryllis T-D Dthes3QThes	Amaryllis T-D-N
Model	25 queries	25 queries	25 queries	25 queries	25 queries
doc=Okapi, query=npn	**45.75**	**45.45**	**44.28**	**44.85**	**53.65**
doc=Lnu, query=ltc	43.07	44.28	41.75	43.45	49.87
doc=dtu, query=dtc	39.09	41.12	40.25	42.81	47.97
doc=atn, query=ntc	42.19	43.83	40.78	43.46	51.44
doc=ltn, query=ntc	39.60	41.14	39.01	40.13	47.50
doc=ntc, query=ntc	28.62	26.87	25.57	26.26	33.89
doc=ltc, query=ltc	33.59	34.09	33.42	33.78	42.47
doc=lnc, query=ltc	37.30	36.77	35.82	36.10	46.09
doc=bnn, query=bnn	20.17	23.97	19.78	23.51	24.72
doc=nnn, query=nnn	13.59	13.05	10.18	12.07	15.94

Table 17a: Average precision of various indexing and searching strategies (Amaryllis)

	Average precision				
Query	Amaryllis T-D	Amaryllis T-D Qthes	Amaryllis T-D Dthes	Amaryllis T-D Dthes3Qthes	Amaryllis T-D-N
Model	25 queries	25 queries	25 queries	25 queries	25 queries
doc=Okapi, query=npn	45.75	45.45	44.28	44.85	53.65
5 docs / 10 terms	47.75	47.29	46.41	46.73	55.80
5 docs / 50 terms	**49.33**	48.27	47.84	47.61	**56.72**
5 docs / 100 terms	49.28	48.53	47.78	47.83	56.71
10 docs / 10 terms	47.71	47.43	46.28	47.21	55.58
10 docs / 50 terms	49.04	48.46	48.49	48.12	56.34
10 docs / 100 terms	48.96	**48.60**	**48.56**	**48.29**	56.34
25 docs / 10 terms	47.07	46.63	45.79	46.77	55.31
25 docs / 50 terms	48.02	47.64	47.23	47.85	55.82
25 docs / 100 terms	48.03	47.78	47.38	47.83	55.80

Table 17b: Average precision using blind-query expansion (Amaryllis)

From the achieved average precision depicted in Tables 17a and 17b, we cannot infer that the available thesaurus is really helpful in improving retrieval effectiveness, at least as implemented in this study.

Run name	Query	Form	Model	Thesaurus	Query expansion	Av. precision
UniNEama1	T-D	automatic	Okapi	no	25 docs / 50 terms	48.02
UniNEama2	T-D	automatic	Okapi	with query terms	25 docs / 25 terms	47.34
UniNEama3	T-D	automatic	Okapi	with documents	25 docs / 50 terms	47.23
UniNEama4	T-D	automatic	Okapi	both query & doc	10 docs / 15 terms	47.78
UniNEamaN1	T-D-N	automatic	Okapi	no	25 docs / 50 terms	55.82

Table 18: Official Amaryllis run descriptions

Conclusion

For our second participation in CLEF retrieval tasks, we suggested a general stopword list and stemming procedure for the French, Italian, German, Spanish and Finnish languages. We also suggested a simple decompounding approach for the German language. For the Dutch, Finnish and German languages we were to consider 5-gram indexing and word-based (and decompounding-based) document representations to be distinct and independent sources of evidence on document content, and it would be a good practice to combine these two (or three) indexing schemes.

To improve bilingual information retrieval, we suggest using not only one but two or three different translation sources to translate the query into the target languages. Such a combination seems to improve the retrieval effectiveness. In the multilingual environment, we demonstrated that a learning scheme such as logistic regression could perform effectively. As a second best solution, we suggested using a simple normalization procedure based on the document score.

Finally, in the Amaryllis experiments, we studied various possible ways we could use a specialized thesaurus to improve average precision. However, the various strategies used in this paper do not demonstrate clear enhancement over a baseline that ignores the term relationships stored in the thesaurus.

Acknowledgments

The author would like to thank C. Buckley from SabIR for giving us the opportunity to use the SMART system, without which this study could not have been conducted. This research was supported in part by the SNSF (Swiss National Science Foundation) under grants 21-58 813.99 and 21-66 742.01.

References

[Buckley 1996] Buckley, C., Singhal, A., Mitra, M. & Salton, G. (1996). New retrieval approaches using SMART. In Proceedings of TREC'4, (pp. 25-48). Gaithersburg: NIST Publication #500-236.

[Chen 2002] Chen, A. (2002). Multilingual information retrieval using English and Chinese queries. In C. Peters, M. Braschler, J. Gonzalo & M. Kluck (Eds.), Evaluation of cross-language information retrieval systems. Lecture Notes in Computer Science #2409. Berlin: Springer-Verlag.

[Figuerola 2002] Figuerola, C.G., Gómez, R. & Zazo Rodríguez, A.F. (2002). Stemming in Spanish: A first approach to its impact on information retrieval. In C. Peters, M. Braschler, J. Gonzalo & M. Kluck (Eds.), Evaluation of cross-language information retrieval systems. Lecture Notes in Computer Science #2409. Berlin: Springer-Verlag.

[Flury 1997] Flury, B. (1997). *A first course in multivariate statistics*. New York: Springer.

[Fox 1990] Fox, C. (1990). A stop list for general text. *ACM-SIGIR Forum*, 24, 19-35.

[Kraaij 1996] Kraaij, W. & Pohlmann, R. (1996). Viewing stemming as recall enhancement. In Proceedings of the 19th International Conference of the ACM-SIGIR'96, (pp. 40-48). New York: The ACM Press.

[Kwok 1995] Kwok, K.L., Grunfeld, L. & Lewis, D.D. (1995). TREC-3 ad-hoc, routing retrieval and thresholding experiments using PIRCS. In Proceedings of TREC'3, (pp. 247-255). Gaithersburg: NIST Publication #500-225.

[Le Calvé 2000] Le Calvé, A., Savoy, J. (2000). Database merging strategy based on logistic regression. *Information Processing & Management*, 36(3), 341-359.

[Lovins 1968] Lovins, J. B. (1968). Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, 11(1), 22-31.

[McNamee 2002] McNamee, P. & Mayfield, J. (2002). JHU/APL Experiments at CLEF: Translation Resources and Score Normalization. In C. Peters, M. Braschler, J. Gonzalo & M. Kluck (Eds.), Evaluation of Cross-Language Information Retrieval Systems. Lecture Notes in Computer Science #2409. Berlin: Springer-Verlag.

[Molina-Salgado 2002] Molina-Salgado, H., Moulinier, I., Knutson, M., Lund, E. & Sekhon, K. (2002). Thomson legal and regulatory at CLEF 2001: Monolingual and bilingual experiments. In C. Peters, M. Braschler, J. Gonzalo & M. Kluck (Eds.), Evaluation of cross-language information retrieval systems. Lecture Notes in Computer Science #2409. Berlin: Springer-Verlag.

[Monz 2002] Monz, C. & de Rijke, M. (2002). The University of Amsterdam at CLEF 2001. In C. Peters, M. Braschler, J. Gonzalo & M. Kluck (Eds.), Evaluation of cross-language information retrieval systems. Lecture Notes in Computer Science #2409. Berlin: Springer-Verlag.

[Porter 1980] Porter, M.F. (1980). An algorithm for suffix stripping. *Program*, 14, 130-137.

[Powell 2000] Powell, A.L., French, J. C., Callan, J., Connell, M. & Viles, C.L. (2000). The impact of database selection on distributed searching. In Proceedings of the 23rd International Conference of the ACM-SIGIR'2000, (pp. 232-239). New York: The ACM Press.

[Robertson 2000] Robertson, S.E., Walker, S. & Beaulieu, M. (2000). Experimentation as a way of life: Okapi at TREC. *Information Processing & Management*, 36(1), 95-108.

[Savoy 1999] Savoy, J. (1999). A stemming procedure and stopword list for general French corpora. *Journal of the American Society for Information Science*, 50(10), 944-952.

[Savoy 2002a] Savoy, J. (2002). Cross-language information retrieval: Experiments based on CLEF-2000 corpora. *Information Processing & Management*, to appear.

[Savoy 2002b] Savoy, J. (2002). Report on CLEF-2001 Experiments: Effective Combined Query-Translation Approach. In C. Peters, M. Braschler, J. Gonzalo & M. Kluck (Eds.), Evaluation of cross-language information retrieval systems. Lecture Notes in Computer Science #2409. Berlin: Springer-Verlag.

[Savoy 2002c] Savoy, J. (2002). Recherche d'informations dans des corpus en langue française : Utilisation du référentiel Amaryllis. *TSI, Technique et Science Informatiques*, 21(3), 345-373.

[Singhal 1999] Singhal, A., Choi, J., Hindle, D., Lewis, D.D. & Pereira, F. (1999). AT&T at TREC-7. In Proceedings TREC-7, (pp. 239-251). Gaithersburg: NIST Publication #500-242.

[Sproat 1992] Sproat, R. (1992). *Morphology and computation.* Cambridge: The MIT Press.

[Voorhees 1995] Voorhees, E.M., Gupta, N.K. & Johnson-Laird, B. (1995). The collection fusion problem. In Proceedings of TREC'3, (pp. 95-104). Gaithersburg: NIST Publication #500-225.