IRENA: Information Retrieval Engine based on Natural language Analysis

A.T. Arampatzis*, C.H.A. Koster, T. Tsoris

University of Nijmegen, CSI, Postbus 9010, 6500 GL Nijmegen, The Netherlands.
Tel: +31 24 3653147, Fax: +31 24 3553450
{avgerino|kees|tsoris}@cs.kun.nl

Abstract

The experimental IRENA system was developed to study the improvement of *precision* and *recall* in document retrieval systems by means of Natural Language Processing (NLP) techniques. The NLP component deals with the *morphological, lexical* and *syntactical* part of the English language. For the purpose of syntactical analysis of both queries and documents, the power of the *AGFL* formalism was explored in describing and developing a syntactical analyzer for the English noun phrase with a large lexicon. The *noun phrase co-occurrence hypothesis* was formulated and tested as a new relevance criterion in achieving high levels of precision. Furthermore, the problem of calculating recall in non-indexed collections was partially solved by introducing a new measure, *relative recall*.

The system was tested on a small corpus of English language documents concerned with pop music. The results of this experiment are reported, and some conclusions drawn on the viability of the techniques.

In the approach taken, all linguistic knowledge is encapsulated in the grammar and the lexicon. The language is merely a parameter. Consequently, the same techniques can also be applied for other, more inflected languages.

Keywords

Information Retrieval, Natural Language Processing, Noun Phrase, Recall, Syntactic Analysis.

*http://www.cs.kun.nl/~avgerino

1. Introduction

The immense increase in the number of electronic documents that reside all over the world and the increasing desire to search and obtain useful information from them has given new impetus to the research for sophisticated Information Retrieval (IR) techniques. Nowadays, users confront large collections of documents that dynamically change day by day. The development of precise and efficient retrieval systems is indispensable.

Our aim is to construct an intelligent and efficient document retrieval system that has the following characteristics:

- achieve high levels of precision,
- be domain independent,
- apply to any collection of full-text documents without pre-processing.

In order to achieve high levels of precision we investigate the use of NLP techniques. Domain independence is accomplished by focusing on syntactic techniques, using domain independent lexica and grammars. We do not primarily pre-process the documents manually. Pre-processing documents linguistically can be time-consuming, especially when the documents are increasing in number and being updated day by day, therefore we shall deal with raw texts. It was decided to improve the recall of the system by means of *query expansion.*

Our research [ARTS96] led to the development of IRENA experimental system. As will be reported, a small-scale experiment has proved the efficiency of IRENA, but many aspects are still under research and the system is continuously being upgraded. In this paper we describe the initial version of IRENA and the directions for future improvements.

2. Architecture of IRENA

IRENA consists of four sub-systems, a *syntactical analyzer*, a *lexical expander*, a *morphological expander* and a *retrieval system.* A brief overview of the components of the IRENA system is given here and the most important of them, such as the syntactical analyzer and the morphological and lexical expander will be described in detail in the following sections.

In order to develop a user-friendly system, IRENA accepts queries in the form of English noun phrases. From each query, certain keywords are extracted by the syntactical analyzer. These are first stemmed and then expanded by the morphological expander, obtaining all lexical and morphological variants of the keywords. Use is made of a database of synonyms to enhance the lexical expansion. The expanded query and the collection of documents are given as input to the retrieval sub-system which simply uses UNIX's `egrep`, which is quite adequate given the size of the collection that was used for the experiment (see section 7.1.) and the fact that our research is concerned with the relatedness of retrieved documents and not with fast retrieval[1]. The documents that `egrep` selects are processed by the syntactical analyzer that analyzes only those document fragments in which keyword co-occurrence appears and takes account of the syntactical relations between the keywords. The output is presented to the user in a ranked fashion, dependent on the morphological, lexical and syntactical relations that have been noticed. The user decides which of the retrieved documents are related to his requests and the effectiveness of the system is estimated based on his judgments. An overview of IRENA is given in figure 1.

[1] `egrep` proved fast enough for our experiment. We have measured a search rate of approximately 3.5 Mb/sec on a SUN SPARCstation.

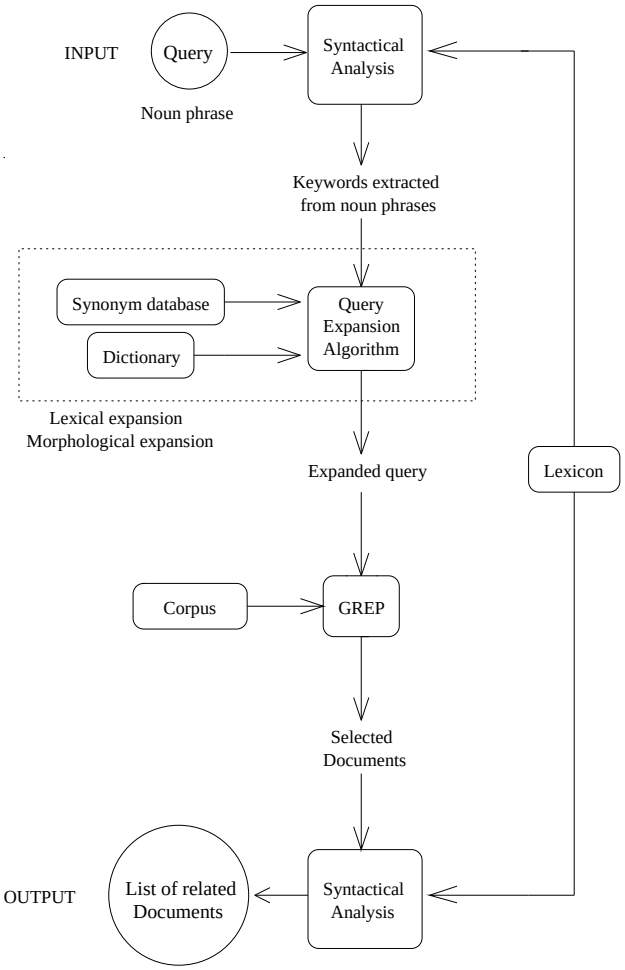

Figure 1: *Overview of the IRENA system*

3. The Syntactical Analyzer

The syntactical analyzer of IRENA was generated automatically from a grammar of English by the AGFL[2] parser generator. The analyzer is capable of recognizing and extracting noun phrases from a document, using a large lexicon.

3.1. The Grammar

The formalism used for the description of the English noun phrase is AGFL (Affix Grammar over a Finite Lattice) [KOST91]. AGFLs are a very simple form of unification grammars, with set-valued features. They are particularly suited to the description of the surface structure of sentences in natural languages.

AGFL grammars have shown their power in the past in describing fragments of several NLs in a useful, reliable, robust and comprehensive way. Although AGFLs can be transformed to CF grammars, they manage to reduce the size of the grammars and make the construction of a complicated grammar for a NL feasible and in a much more understandable way.

Rather than using a large, linguistically motivated grammar, in the experiment we started out from a small AGFL grammar that deals satisfactorily with the English noun phrase

[2]For more information on the AGFL system, see: `http://www.cs.kun.nl/agfl`

[MEKO62], which was adapted to the intended Information Retrieval application. The main revisions can be outlined as follows:

- **Wild-card parsing** is the parsing of selected segments of a sentence. The segments we are interested in are the noun phrases. The selection is expressed by the grammatical rules. The syntactical analyzer that was created is able to capture and analyze the longest noun phrase in a sentence (*latest closure*) and discards the rest of the words as garbage. The garbage is explicitly described in the grammar as a chain of characters.

- **Proper Name** recognition is an important feature for Information Retrieval applications, that had to be embedded in the syntactical analyzer, as in Information Retrieval the documents may contain a plethora of proper names. The method of inserting proper names in the lexicon is inadequate because it implies a pre-processing of the documents to collect all the proper names. Consequently, syntactic rules describing the proper names were constructed, under the premise that each word that does not occur in the lexicon and starts with a capital letter is a proper name. Although this is an ad hoc assumption, it is true for the majority of proper names, with few exceptions. The syntax rule for proper names is capable to recognize:

  1. multi-word proper names (e.g. *Red Hot Chili Peppers*),
  2. single or double quoted proper names (e.g. *"The Smiths"*, *'The Cure'*),
  3. abbreviations (e.g. *U.S.A.*, *R.E.M.*) and
  4. proper names which contain special characters (e.g. *Roger O' Donnell*, *J & B*, *Ice-T*, etc.).

The syntactical analyzer generated deals quite efficiently with ambiguity. In natural language parsing, the major problem we have to deal with, is the problem of *structural ambiguity* and, as a result, the exponential increase of the number and time of analyses. The method that was followed is syntactic *under-specification*. This method constrains the parsing process by incorporating restrictive rules in the grammar [BAOL95]. The rules were formed in such a way that the parser analyzes the longest noun phrase in a sentence and does not investigate the other alternative rules of the grammar.

3.2. The Lexicon System

For the experiment we made use of a lexicon which was constructed from the well-known WordNet[3] lexicon [5PWN93]. We preferred the use of an external lexicon, rather than the insertion of actual words in the form of terminal productions into the grammar, for two reasons: parsing efficiency and abstraction. Otherwise, the analyzer becomes extremely large in size and poor in speed, as it matches every alternative serially. Moreover, maintenance of the grammar becomes unmanageable.

The system that was used for the creation of the lexicon is *LEXGEN*. Words are attributed to several grammatical categories (e.g. nouns, adverbs, adjectives, etc.). The lexicon file has a special data structure that allows compression of data and fast matching. In order to treat *lexical ambiguity*, we have tried to create a lexicon in which each word belongs to as few as possible lexical categories. Adjectival participles have been kept apart from adjectives and

[3]Created by Cognitive Science Laboratory, Princeton University, 221 Nassau St. , Princeton, NJ 08542. WordNet is available for anonymous ftp from `clarity.Princeton.edu` and `ftp.ims.uni-Stuttgart.de` .

exist only as participles (present or past). The same has been done in the case of gerunds which have been removed from nouns. Nevertheless, lexical ambiguity occurs in several syntactical analyses. The number of word-forms that constitute the lexicon are summarized in Table 1.

Part of speech	**Words**
Nouns	173,000
Adjectives	16,251
Adverbs	3,553
Present participles	1,022
Past participles	2,104
Prepositions	161
Pronouns	33
Determiners	12
Articles	3
Total	196,139

Table 1: *Words of the English Lexicon*

The result of this interfacing is a high-speed, easily-maintainable and expandable noun-phrase syntactical analyzer for the English noun phrase. To measure the speed of the analyzer we have parsed 100 sentences, taken from the Unix manuals, with a mean number of 8.61 words per sentence. The parsing speed was 83.2 words/sec on a SPARCstation.

3.3. Parsing

Each submitted query had to be syntactically analyzed in order to extract the useful information. For this reason, each query is parsed and from the resulting parse trees keywords are extracted. An English noun phrase does not contain only nouns but also adjectives, articles, prepositions, adverbs and proper names (which are nouns but are considered as a separate category due to their importance in IR). Nouns, proper names and adjectives are the only pieces considered useful. Articles and prepositions are redundant, but adverbs can sometimes give extra information.

The extraction of keywords is performed by scanning the output of the analyzer and by submitting only the nouns, adjectives, present participles, past participles and proper names to the query expansion sub-system. If any lexical ambiguity occurs, all possible parts-of-speech are taken into account. For example, the query `Q: female vocalists` has the two syntactical analyses given in Figure 2. On the basis of these analyses, "female" and "vocalists" are given to the expansion sub-system, "vocalists" as noun, but "female" is tagged both as noun and adjective.

For the analysis of the documents, a sentence boundary recognizer was constructed which splits free text into sentences by a number of heuristic rules. This is an important issue, because the analyzer receives only sentences as input.

```
parsing 1

Sentence
  noun phrase(PLUR, THIRD, NOM|DAT|ACC)
    noun part(PLUR, THIRD, NOM|DAT|ACC)
      noun group(PLUR, NOM|DAT|ACC)
        adjective phrase
          ADJE(ABSO)
            "female"
        noun group(PLUR, NOM|DAT|ACC)
          NOUN(PLUR, NOM|DAT|ACC)
            "vocalists"


parsing 2

Sentence
  noun phrase(PLUR, THIRD, NOM|DAT|ACC)
    noun part(PLUR, THIRD, NOM|DAT|ACC)
      noun group(PLUR, NOM|DAT|ACC)
        NOUN(SING, NOM)
          "female"
        NOUN(PLUR, NOM|DAT|ACC)
          "vocalists"
```

Figure 2: *Analyses of the query Q: female vocalists*

4. Lexical Expansion

Different words that share the same general meaning (*synonyms*) have to be considered during query expansion [GREF92]. It is very fortunate that information about synonyms can be obtained from most modern on-line lexical databases, such as WordNet.

The most obvious difference between WordNet and a standard dictionary is that WordNet divides the lexicon into four syntactic categories (nouns, verbs, adjectives, adverbs), a feature which can be systematically exploited in NLP. *Collocations* are also included in WordNet and this could be very useful in some cases which synonyms are obtained only indirectly by putting two or more words together. An example of such periphrastic synonyms can be given for "biography".

{ biography, life, life_story, life_history }
`(an account of the series of events making up a person's life)`

In IRENA, each keyword and its part-of-speech is submitted to WordNet, whereupon nouns are expanded with noun synonyms and adjectives with adjectival synonyms. Present participles that function as nouns (gerunds) and adjectival participles brought up a problem. Gerunds and adjectival participles are recognized by the parser simply as participles and it is not clear whether they function as nouns or as adjectives. WordNet's lexical categories

of nouns and adjectives also contain the most used gerunds and adjectival participles. It was decided that present participles would be expanded both as nouns and as adjectives and past participles as adjectives. We found out that this was an effective and fast solution, as erroneous expansions of gerunds with adjectival synonyms resulted only in an insubstantial, almost insignificant, loss of precision. A rather large number of proper names exist in WordNet in the nouns category and as a consequence, proper names also are expanded. This is considered extremely powerful; consider as an example the expansion of "USA":

> **{ United_States, United_States_of_America, America, US, U.S., USA, U.S.A }**
> `(there are 50 states in the US)`

The synonyms of a word depend on the meaning of the word in a specific context. For instance, when the word "note" is used in music contexts, the word "tune" can be considered as a similar word, but "comment" cannot. It is obvious that, disregarding word meanings, the expansion is not reliable. In WordNet, nouns, verbs, adjectives and adverbs are organized into synonym sets – lists of synonymous word forms that are interchangeable in some context. Combining all synonym sets for the keywords and disregarding the word meanings, each keyword of the query `Q: popular bands` is expanded as following:

> **popular**, `demotic, lay, plain, nontechnical, unspecialized, untechnical, pop.`
>
> **band**, `set, circle, lot, stria, striation, banding, stripe, dance band, dance orchestra, frequency band, ring.`

Words like "nontechnical" and "lay" are not synonyms of "popular" in music contexts, as well as "stripe" and "ring" are not synonyms of "band". Fortunately, this has almost no impact on precision and recall, but only on the retrieval speed due to the needless searches for those synonyms. This is true only when the corpus is related to one general subject (i.e. music or medicine). The words "demotic" and "stria" hardly occur in music contexts, though the unfortunate co-occurrence of "lay" and "lot" will result in some loss of precision, but this is also rare. The question is semantical rather than lexical. We experimented with the solution of showing each synonym set to the user, and asking for a confirmation of its relevance before using it in the lexical expansion. As an example, a user who submits a query like the preceding will reject synonym sets like:

> **{ band, stria, striation }**
> `(a stripe of contrasting color; ''chromosomes exhibit characteristic bands'')`
>
> **{ band, frequency band }**
> `(band of radio frequencies for e.g. transmitting a TV signal)`

and will accept the synonym set:

> **{ dance band, band, dance orchestra }**
> `(a group of musicians playing popular music for dancing)`

By doing that, the ideal expansion is obtained:

> **popular**, `pop.`
>
> **band**, `dance band, dance orchestra.`

The users (students of Computer and Cognitive Science) were not very satisfied with this solution — too much interaction was needed. Considering a particular subject of a corpus and the psychological profile[4] of a user, this expansion can be automatically achieved in future versions of IRENA.

5. Morphological Expansion

After the lexical expansion, morphological expansion is applied to every keyword and to its synonyms to obtain all morphological variants.

Morphology is the area of linguistics concerned with the internal structure of words and is usually is divided into two subclasses, *inflectional* and *derivational* morphology. Information retrieval has generally not paid much attention to word structure, other than to account for some of the variability in word forms via the use of *stemmers* [CROF94]. *Stemming* is any process that strips the suffixes from a word to obtain the root word. Inflectional morphology describes the predictable changes a word undergoes as a result of syntax. The most common changes are:

- The plural and the possessive form for nouns.
- The comparative and superlative form for adjectives.
- The past tense, past participle and progressive form for verbs.

These changes have no effect on a word's part-of-speech; a noun still remains a noun after pluralization. Inflectional morphological variations always occur after derivational forms. Inflectional variation is typically associated with syntax, and has relatively little impact on a word's part-of-speech and meaning; derivational morphology may or may not affect a word's part-of-speech or meaning. As the meaning of a word may alter, the issue is rather semantical and we decided not to deal with derivational morphology at present. A likely approach would be to take into account only those derivational variants whose meaning is related to the root word. It is not so difficult to decide which variants are related to the root word due to the fact that dictionaries usually list a word-form separately, if it has a meaning that is distinct from that of the root.

In IRENA, inflectional expansion is applied to nouns, proper names and adjectives, after each word has been stemmed, because the expander assumes root words as input. Nouns are converted into singular nominative form and adjectives in comparative or superlative form into the base form. The stemmer which is used is quite similar to the *revised Porter* stemmer [KROV93], a modification of the Porter algorithm that checks a word against the dictionary after each deduction step. This prevents "calories" from being converted to "calory". The algorithm uses exception lists for each syntactic category due to irregularity in inflection of some words, consequently, "wolves" and "best" are stemmed to "wolf" and "good" after a look-up in the exception lists of nouns and adjectives before the deduction process.

The expander conflates the singular and the possessive forms of nouns, and the comparative and superlative form of adjectives. A list of the gradable adjectives is used for this purpose. As some nouns have irregular plural form, they must be checked in the exception list first. If the noun is not irregular, the English grammar rules [ALEX88] are followed for the creation of the plural form, otherwise the plural is taken directly from the exception list. For the conflation of the genitive, the rules [ALEX88] applied to the singular as well as the

[4] This is important for resolving jargon semantical ambiguities

plural form of the noun. All common nouns in English fall into one of two sub-classes: they may be either countable or uncountable and that distinction is fundamental for the existence of the plural. Unfortunately, strict classifications of nouns are in many cases unreliable, as some nouns which are normally uncountable can be used as countable in certain contexts. For instance, the noun "weather" is normally uncountable, but it can be said "I go out all weathers". The distinction of nouns in countable and uncountable is not taken into consideration, so some nouns may be expanded into nonexistent plural forms. This has a slight negative impact on the retrieval speed, due to the useless searches for nonexistent words.

Adjectival participles and gerunds are not stemmed and expanded at all, due to the fact that adjectival participles are not inflected and gerunds have no plural. Proper names are treated in a different way. They are not stemmed and the expander conflates only the genitive in the given number.

6. Retrieval Strategy

An ideal retrieval strategy would be based on some measure of the "nearness" of one noun phrase (in the query) to another (in the document). Although similar measures had been developed (*logical "nearness"* in [BRU93] and [BRU94]), we investigated in IRENA other, more heuristic strategies that fit for the *Noun Phrase Co-occurrence Hypothesis*.

6.1. The Noun Phrase Co-occurrence Hypothesis

Our basic premise is that words occurring in the same noun phrase tend to share some semantical relation. If two or more nouns and their respective adjectives are found in a single noun phrase, then we can assume that these nouns share some relatedness, even without knowing what they stand for. For example in the phrase

> *... tracks were recorded at the BBC studios for later radio broadcast ...*

the nouns "radio", "broadcast" and the proper name "BBC" which reside in the same noun phrase of the sentence are semantically related. Therefore, searching for the programs of the BBC radio station with the query `Q: radio programs on BBC`, we can retrieve documents containing phrases like the one above and not documents with other forms of co-occurrence like:

> *The transmission of his first radio programs resembled the early years of the creation of BBC empire which ...*
>
> *Ten musicians from the BBC Symphony Orchestra were interviewed in several radio programs of L.A. stations ...*

These phrases are rejected due to the syntactic information that the three words of the query reside in different noun phrases. The last real cases clearly show that extra linguistic processing is more beneficial compared to a *proximity search* that requires words in the user's query to be close to each other in the document.

Of course we can expect some exceptions which do not conform to this hypothesis. We encountered some phrases during the execution of the experiment where the terms of a query existed in one single noun phrase of a document's sentence, but they were not semantically related. In these few cases, the need of a semantic analysis system becomes apparent. A characteristic phrase of this kind and its respective query follow.

`Query: soundtracks of films`

Text: *...In this album, there is a good background, but there is something missing. Either a solo voice or instrument. Or at least a film. Soundtrack without pictures so to speak. ...*

Searching for film soundtracks in general we came across this text. The noun "picture" is a synonym of the noun "film" and belongs to the same noun phrase as "soundtrack". But the meaning of the last sentence is merely that this album could be a soundtrack of a movie but it was not. Notice that the prepositions (`of/without`) have not been taken into account.

6.2. Ranking

The retrieval sub-system of IRENA returns those documents in which at least one variant (lexical, morphological or the keyword itself) of each initial keyword appears. The output of retrieval strategy depends both on the type of variants found and the *distance* between them.

The distance is calculated in text lines. The parameter W indicates the text window size in which any co-occurrence exists. W is increasing and the output is presented in that order. Additionally, the documents in which co-occurrence exists in the smallest window size are syntactically analyzed in order to check if the NP co-occurrence hypothesis is satisfied. Based on the assumption that an English NP is not more that two text lines, W_1 is in this case defined as 2.

The importance (in descending order) of co-occurrence categories were determined as: initial keyword, morphological variant, synonym co-occurrence. This is the right order according to a small-scale experiment, since following this order the precision decreases. Summarizing, IRENA ranks the output using the following order:

(highest weight)
keyword co-occurrence in NP
morphological variant co-occurrence in NP
synonym co-occurrence in NP
keyword co-occurrence in W_1
morphological variant co-occurrence in W_1
synonym co-occurrence in W_1
$\vdots$
keyword co-occurrence in W_n
morphological variant co-occurrence in W_n
synonym co-occurrence in W_n
(lowest weight)

7. The Experiment

The ideas that are outlined in the next sections were implemented and tested in a small-scale experiment.

7.1. The Corpus

The corpus which was used for the small-scale experiment is a set of documents about music (e.g. magazine articles, Frequently Asked Questions about artists and groups, interviews,

album reviews, etc.) collected from the Internet. Some statistics for this corpus are shown in Table 2. The documents are unconstrained Internet prose, full of colloquialisms, misspellings

	Corpus
Number of documents	633
Mean words per document	1695
Number of words in collection	1,072,762
Total size	6,752 Kb

Table 2: *Statistics on text corpus*

and ungrammatical statements. Many of them refer to more than one event, artist or kind of music and that makes it hard to categorize them under only one sub-subject of music.

Unique characteristics of this corpus are the large number of proper names and the presence of informal language. Group and artists' names, abbreviations etc. are an interesting problem to elaborate. Furthermore, interviews contain a lot of musical jargon and all these magnify the *lexical* and *semantical ambiguity*. This corpus is chosen for the above characteristics, as well as for its absorbing subject that can seduce the users into formulating queries.

7.2. Quality Measures

Since the particular subject of any document in the collection is unknown (because there is no indexing), a quite significant problem had to be faced very early in calculating the recall. We chose to introduce a new, approximative measure for recall.

First, some standard measures widely used in IR are briefly defined in order to make clear the need to redefine recall.

Standard Measures

Let the set O be the corpus and a query q retrieves the subset B_q while A_q was intended. If the output of the retrieval strategy depends on a parameter λ such as the co-ordination level or the distance between the keywords found, precision and recall of a system for q are customarily defined as:

$$Precision_\lambda(q) = \frac{|A_q \cap B_{q\lambda}|}{|B_{q\lambda}|} \tag{1}$$

$$Recall_\lambda(q) = \frac{|A_q \cap B_{q\lambda}|}{|A_q|} \tag{2}$$

where $|X|$ denotes the number of documents in set X. If Q is a set of requests then the average precision and recall of a system, using *micro-evaluation* [RIJS79] as an averaging technique, can be calculated as follows:

$$Precision_\lambda = \sum_{q \in Q} \frac{|A_q \cap B_{q\lambda}|}{|\widetilde{B_\lambda}|} \tag{3}$$

$$Recall_\lambda = \sum_{q \in Q} \frac{|A_q \cap B_{q\lambda}|}{|\widetilde{A}|} \tag{4}$$

where $|\widetilde{A}|$ and $|\widetilde{B_\lambda}|$ are:

$$|\widetilde{A}| = \sum_{q \in Q} |A_q| \tag{5}$$

$$|\widetilde{B_\lambda}| = \sum_{q \in Q} |B_{q\lambda}|$$

For further discussion about these quality measures and averaging techniques we refer to [RIJS79].

Redefinition of Recall

$|A_q \cap B_{q\lambda}|$ can be calculated simply by users counting the relevant documents of the output. A_q can not be determined and that creates the inability of calculating $Recall_\lambda(q)$. The determination of A_q presupposes reading of each individual document by a human and indexing of it by the subject(s), which is a laborious and time-consuming task. Because of that, we have tried to give a new definition of recall that is easy to calculate under the conditions of the experiment and is as close as possible to the real recall of the system.

Based on the assumption that all the collection is in some respects relevant to every query, we can derive from equation 2, a new measure that we will refer to by the name *Relative Recall* ($RR_\lambda(q)$).

$$RR_\lambda(q) = \frac{|A_q \cap B_{q\lambda}|}{|O|}$$

This new measure does not give reliable results for individual queries, but it can be still used for comparing recall between two or more queries. Let's examine some conditions under which results will be close to reality.

If n is the number of queries that have been submitted to the system, average RR_λ can be defined from (4), as below.

$$RR_\lambda = \sum_{i=1}^{n} \frac{|A_{q_i} \cap B_{q_i\lambda}|}{|\widetilde{A}|} = \frac{\sum_{i=1}^{n} |A_{q_i} \cap B_{q_i\lambda}|}{n|O|}$$

$RR(q)$ is usually much smaller than $Recall(q)$ due to the impossibility of initial assumption, so the distribution of RR_λ values is not regular in $[\,0\,,\,1]$ but values tend to 0. To force the distribution of the values in all the interval, RR_λ is normalized by a new factor.

$$RR_\lambda = \frac{|O|}{\max_{1 \le i \le n}(|A_{q_i} \cap B_{q_i}|)} \frac{\sum_{i=1}^{n} |A_{q_i} \cap B_{q_i\lambda}|}{n|O|} \Rightarrow$$

$$\Rightarrow RR_\lambda = \frac{\sum_{i=1}^{n} |A_{q_i} \cap B_{q_i\lambda}|}{n \ \max_{1 \le i \le n}(|A_{q_i} \cap B_{q_i\lambda}|)} \tag{6}$$

We will prove that, under some conditions, RR_λ values normalized by this factor are very close to real $Recall_\lambda$.

If we assume that all $|A_{q_i}|$ values are close to the average of the relevant documents per query[5], denoted by constant c, that is $|A_{q_1}| \approx |A_{q_2}| \approx \ldots \approx |A_{q_n}| \approx c$, we will get from (4) the following:

$$Recall_\lambda = \sum_{i=1}^{n} \frac{|A_{q_i} \cap B_{q_i\lambda}|}{|\widetilde{A}|} =$$

[5] means that $Gen(q_i)$ is approximately constant for all i. For the definition of *Generality* refer to [RIJS79]. In order to check how realistic is that assumption, analyses on "established" IR test-collections must be done

$$= \sum_{i=1}^{n} \frac{|A_{q_i} \cap B_{q_i\lambda}|}{\sum_{i=1}^{n} |A_{q_i}|} \approx \sum_{i=1}^{n} \frac{|A_{q_i} \cap B_{q_i\lambda}|}{n\,c} \Rightarrow$$

$$\Rightarrow Recall_\lambda \approx \sum_{i=1}^{n} \frac{|A_{q_i} \cap B_{q_i\lambda}|}{n\,c} \tag{7}$$

It is quite possible for the system to succeed to retrieve all the relevant documents in at least one query q_+, that means $A_{q_+} \cap B_{q_+\lambda} = A_{q_+}$ and $|A_{q_+}| \approx c$.

Probably, $|A_{q_+}| \geq |A_{q_i} \cap B_{q_i\lambda}| \quad , i = 1, 2, \ldots, n$ and consequently:

$$c \approx |A_{q_+}| = \max_{1 \leq i \leq n} (|A_{q_i} \cap B_{q_i\lambda}|) \tag{8}$$

From equations (7), (8) and (6) we conclude,

$$Recall_\lambda \approx \frac{\sum_{i=1}^{n} |A_{q_i} \cap B_{q_i\lambda}|}{nc} \approx \frac{\sum_{i=1}^{n} |A_{q_i} \cap B_{q_i\lambda}|}{n \max_{1 \leq i \leq n} (|A_{q_i} \cap B_{q_i\lambda}|)} = RR_\lambda \tag{9}$$

In the experiment, we used equation 6 in order to calculate the average recall values of the system. When all the requests to the system had been completed, we realized that for a subset of queries $Q, Gen(q), q \in Q$ was not so close to the average Gen. That suppressed RR_λ values in the range 0%-35%, so we decided to divide all RR_λ values by the maximum RR_λ for reasons of better representation of the Precision-Recall (P-R) curves in section 7.3.. The reader must keep this in mind, in order to understand the recall values close to 100% that occur in some results.

7.3. Performance

Forty-four noun phrase queries were submitted to the system and an average of 2.6 keywords per query were extracted. The expansion with synonyms resulted in 4.1 more keywords, that is about 1.6 synonyms per initial keyword. The morphological expansion of all keywords and their synonyms added an average of 14.2 more search words. Consequently, for every query, an average of 20.9 keywords were submitted to the retrieval sub-system. The precision-recall results are summarized in table 3 and graphically in figure 3. The initial *keyword based search*

	Keyword		**Morphological**		**Lexical**	
Window	Precision %	Recall %	Precision %	Recall %	Precision %	Recall %
NP	100.00	6.31	95.65	23.14	91.38	27.90
2	79.17	19.98	76.45	49.44	71.61	58.41
3	74.60	24.75	70.62	58.41	66.33	69.43
4	72.97	28.40	68.05	66.83	63.40	78.40
6	70.21	34.71	65.22	76.79	59.59	91.58
8	69.44	39.48	63.87	84.22	56.38	100.00

Table 3: *Precision-Recall results*

is compared to *morphological search* (search for initial keywords and their morphological variants) and *lexical search* (search for initial keywords, their synonyms, and morphological

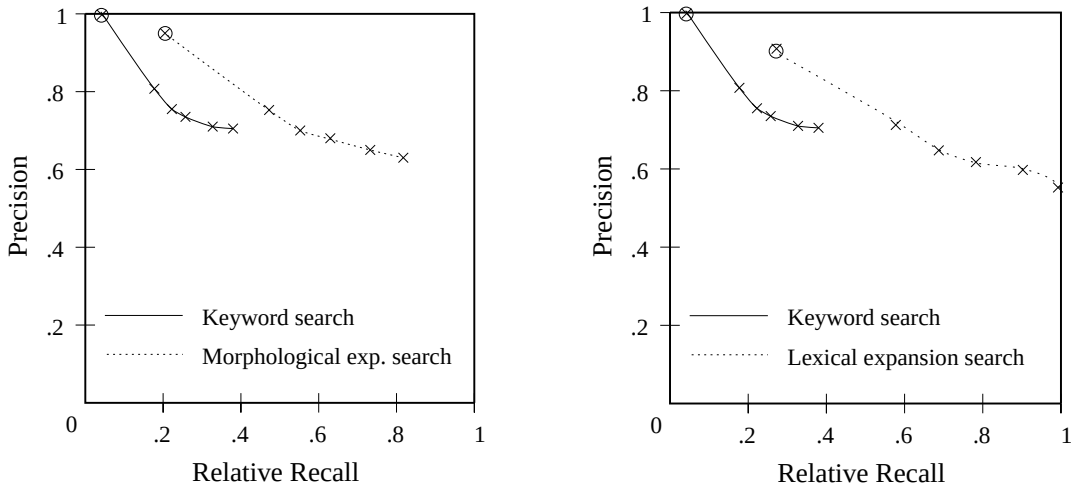

Figure 3: *Precision-Recall graphs of keyword/morphological/lexical search*

variants of these). Also, the simple co-occurrence in 2-8 text lines is compared to NP co-occurrence. In figure 3, the points which are included in circles represent precision and recall for NP co-occurrence.

The use of NP co-occurrence resulted in very high precision levels, above 90%, in all search types especially in keyword search; in this case, precision was even 100%, possibly because of a tendency of the user and formulator of the queries to think primarily in terms of keywords, rather than phrases. Conversely, recall with NP co-occurrence was extremely low compared to simple co-occurrence in a text window. As window size increases from 2 to 8 lines, it seems that better recall is gained at the price of a slight drop in precision. However, it is found that by increasing the size to more than 16 lines, precision is dramatically lowered to 25-35%. Upon enlarging the window, keywords may appear in different paragraphs with possibly different subjects, which accounts for this large drop in precision. A window size of 4 to 8 lines gives reasonable levels of precision and recall.

Expanding queries with lexical and morphological variants led to a remarkable increment in recall, up to 60%. The decrement of precision, which in the worst case was 13%, can be considered as insubstantial compared to the recall gained.

It should be realized that these experimental results are quite tentative, the queries used in the experiment may not be considered as representative and a better controlled experiment with "average" users still has to take place. Still, the experiment has been very conclusive to us in pointing out what directions to pursue.

8. Conclusions and Future Improvements

The small-scale experiment in IRENA has (again) proved that lexical and morphological expansion of queries is indispensable for high recall and results in an insubstantial average loss of precision, hence is highly recommended. This holds in spite of the fact that the natural language used was English, which is weak in morphology and poor in syntax. Experiment has to show whether this also holds for highly inflected languages like e.g. modern Greek.

The NP co-occurrence criterion has proved to be successful in determining whether keywords are semantically related and achieves a much better precision that *proximity search.*

The low recall obtained suggests the generalization of the NP hypothesis to wider classes of phrases to delimit the semantic relatedness between words (verbal phrases, anaphora).

At any rate, the NP co-occurrence criterion can also be used in the future for relevance feedback.

The dramatically low recall achieved could be interpreted in two different ways: One could argue (like [SMEA92] and [CRGA90]) that use of the noun phrase shows no promise in improving the performance of IR systems. We argue, on the other hand, that we should retain the noun phrase as a unit of co-occurence, but should investigate the possibilities of enhancing the recall without loosing too much precision.

A number of ideas merit investigation:

1. the treatment of *anaphora* in order to catch references to previous noun phrases, and
2. the possibility to apply *syntactic normalization* in order to deal with the rich choice of alternative syntactic formulation for one same noun phrase , such as: *air pollution, polluted air, pollution of the air, air is polluted*, etc.

Based on the experience with IRENA, the noun phrase hypothesis and the effects of anaphora resolution and syntactic normalization are presently being investigated in two projects:

1. the Information Filtering projrect PROFILE[6] , at the University of Nijmegen, the effect of syntactic normalization will be investigated in the context of English documents.
2. the DoRo project (ESPRIT HPC), which aims at the development of a system for the automatic classification and routing of full-text documents.

At subsequent conferences we hope to inform you of the results.

[6]For more information see: `http://hwr.nici.kun.nl/~profile` .

References

[5PWN93] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. *"Five Papers on WordNet"*, Cognitive Science Laboratory, Princeton University, August 1993. Available for anonymous ftp from `ftp://clarity.princeton.edu/pub/wordnet/5paper.ps`

[ALEX88] L.G. Alexander. *Longman English Grammar*, Longman Group UK Limited, 1988.

[ARTS96] A.T. Arampatzis and Th. Tsoris. *"A Linguistic Approach in Information Retrieval"*, Master's thesis, University of Nijmegen and University of Patras, 1996.

[BAOL95] B. van Bakel and E. Oltmans. *"A Modular Approach to Handling Syntactic Ambiguity"*, Department of Language & Speech and Department of Computer Science, University of Nijmegen, 1995.

[BRUZ93] P.D. Bruza. *"Stratified Information Disclosure - a synthesis between hypermedia and information retrieval"*, PhD Thesis, University of Nijmegen, 1993.

[BRUZ94] P.D. Bruza and J.J. Ijdens. "Efficient Probabilistic Inference through Index Expression Belief Networks", *Proceedings of AI '94, the Seventh Australian Joint Conference on Artificial Intelligence*, World Scientific Publishers, pages 592-599.

[CACR93] J.P. Callan and W.B. Croft. "An Evaluation of Query Processing Strategies using the TIPSTER Collection", *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 347-356.

[CRGA90] W.B. Croft and L. Gay. *"Interpreting Nominal Compounds for Information Retrieval"*, Information Processing and Management, 2691), 21-38, 1990.

[CROF94] J. Xu and W.B. Croft. "Corpus-Specific Stemming using Word Form Co-occurrence", *UMass Technical Report*, University of Massachusetts.

[GREF92] G. Grefenstette. "Use of Syntactic Context to Produce Term Association Lists for Text Retrieval", *15th Ann Int'l SIGIR '92/Denmark-6/92*, pages 89-97.

[KOMO92] H.V. Kontouli and M.-A.G. Mountzia. *"Morphological and syntactical description of the Greek language"*, Master's thesis, University of Nijmegen and University of Patras, 1992.

[KOND92] C.H.A. Koster, M.J. Nederhof, C. Dekkers and A. van Zwol. *"Manual for the Grammar WorkBench"*, Department of Informatics, University of Nijmegen, 1992, pages 11-17.

[KOST91] C.H.A. Koster. "Affix Grammars for Natural Languages", *Attribute grammars, applications and systems*, International summer school SAGA. Lecture notes in computer science 545. Prague, 1991. Springer-Verlag. 358-373.

[KROV93] R. Krovetz. "Viewing Morphology as an Inference Process", *ACM-SIGIR '93-6/93/Pittsburgh, PA, USA*, pages 191-202.

[KUPI93] J. Kupiec. MURAX: "A Robust Linguistic Approach For Question Answering Using An On-Line Encyclopedia", *ACM-SIGIR '93-6/93/Pittsburgh, PA, USA*, pages 181-190.

[MEKO62] L.G.L.Th. Meertens and C.H.A. Koster. *"An Affix grammar for a part of the English language"*, presented at the Euratom Colloquium, University of Amsterdam, 1962.

[NEKO92] M.-J. Nederhof and C.H.A. Koster. "English Language Corpora: Design, Analysis and Exploitation", *Papers from the thirteenth International Conference on English Language Research on Computerized Corpora*, Nijmegen 1992, pages 166-168.

[RIJS79] C.J. van Rijsbergen. *"Information Retrieval"*, Butterworths, second edition, 1979. Also accessible via WWW in: `http:// www.dcs.glasgow.ac.uk/ Keith/Preface.html` .

[SMEA92] A.F. Smeaton. *"Progress in the Application of NLP to Information Retrieval Tasks"*, The Computer Journal, 26(3), 268-278, 1992.