

Head/Modifier Frames for Information Retrieval

C.H.A. Koster and T. Verhoeven

Computing Science Institute,

University of Nijmegen,

The Netherlands,

E-mail: {kees,tjarkv}@cs.kun.nl

Abstract

In this paper we describe a principled method for representing documents by phrases abstracted into Head/Modifier frames. We first discuss the notion of aboutness and the characterization of a full-text document by HM frames. Based on linguistic arguments, a taxonomy of HM frames is derived. We briefly describe the EP4IR parser/transducer of English, present some experimental results concerning the distribution of HM frames in newspaper text and measure the accuracy of the parser, using a method based on the HM frames generated.

1 Introduction

The Information Retrieval community has long has high hopes concerning the value of linguistic techniques. A quote from (van Rijsbergen, 1979):

The time is ripe for another attempt at using natural language to represent documents inside a computer. There is reason for optimism now that a lot more is known about the syntax and semantics of language.

However, the improvement in precision and/or recall expected from the use of phrases in retrieval and text classification has repeatedly been found disappointing.

Although the use of simple noun phrases as indexing terms is now commonly accepted, practical Information Retrieval systems using phrases like the CLARIT system (Evans et al., 1993) do not appear to perform consistently better than those based on keywords. There is a growing conviction that the value of Natural Language Processing to IR is dubious, even among people who tried hard to make linguistically-based IR work (Lewis, 1992; Smeaton, 1997). The predominant feeling, as voiced in (Sparck Jones, 1998), is that only 'shallow' linguistic techniques like the use of stoplists and lemmatization are of any use to IR, the rest is a question of using the right statistics.

In spite of these negative experiences, we are trying to improve the accuracy of automatic document classification techniques by using Linguistically Motivated Indexterms. Our approach differs from previous approaches in important respects:

- we work in the area of text classification rather than that of "classical" IR, so the human vagaries in formulating queries play no role
- we base our work on a principled analysis of phrase representations, as presented here
- unlike many approaches we take not only NP's but also VP's (or rather the frames derived from them) as terms
- we use a parser based on a grammar (rather than a "shallow parser" based on tagging and word patterns) to isolate the phrases
- we will use morphological as well as syntactic normalization and semantic fuzzy matching (as described in (Koster et al., 1999)) to enhance recall.

We hope that in combination these techniques will achieve a significant improvement in document classification.

In this paper we shall first discuss the notion of *aboutness*, which plays a central role in Information Retrieval. We then introduce Head/Modifier (HM) frames as an abstraction of phrases preserving their aboutness, and derive a correspondence between certain important intra-sentence relations and the corresponding frames. Finally we give some experimental results about the distribution of HM frames, related to the accuracy of the parser/transducer used.

2 Aboutness

The notion of *aboutness* is highly central to Information Retrieval (Bruza and Huibers, 1996): the user of a retrieval system expects the system, in response to a query, to supply a list of documents which are *about* that query.

For the purpose of search, the documents are represented ("characterized") by certain terms and the query also has to be formulated in such terms. The set of terms used to be chosen from a fixed thesaurus (in the case of *assigned* terms), but in modern full-text Information Retrieval, exemplified by search engines on the web, it is customary to simply use all words from the document as *extracted* terms (possibly after applying a stop list and lemmatization).

A model-theoretic basis for the notion of aboutness was described in (Bruza and Huibers, 1994):

An information carrier i will be said to be *about* information carrier j if the information borne by j holds in i

which allows the formal reasoning about aboutness. The rather abstract sounding notion of “information carrier” can denote a single term, but also a composition of terms into a structured query or document.

In other retrieval models (Boolean, vector space, logic-based or probabilistic) the notion of aboutness can be defined analogously (see (Bruza and Huibers, 1996)).

Retrieval systems using words as terms are based on an extremely simple form of aboutness:

If the word x occurs in the document then the document is *about* x .

This notion can be refined by introducing a measure for the *similarity* between the query and the document.

We are interested in the use of Linguistically Motivated IndexTerms (LMI, also known a LMT) rather than keywords, for which aboutness can be defined in the same way:

If the phrase x occurs in the document then the document is *about* x .

This approach may still be rather naive, but intuitively it seems clear that phrases provide a more informative document representation than keywords.

2.1 Phrases as indexing terms

The use of linguistically motivated indexing terms has over a long period fascinated many researchers in IR (for an overview see (Sparck Jones, 1999)). Even the particular choice of HM frames as indexing terms is not new: it has been made before by researchers like (Fagan, 1988; Strzalkowski, 1995).

In particular the Noun Phrase enjoys popularity as LMI, because on the one hand it obviously carries a lot of information, and on the other hand it is relatively easy to extract from a document.

2.1.1 Noun Phrases as indexing terms

According to (Winograd, 1983), a Noun Phrase is to be considered as a *reference to or description of a complicated concept*. It is interesting to note that the preferred form of informative titles of articles in the exact sciences appears to be a (complicated) noun phrase. As a query, a noun phrase is definitely more precise than the bag of its constituent words.

The use of simple NP’s as indexing terms is now common practice in many Information Retrieval systems. NP’s can be extracted from a

text using a shallow parser (e.g. (Daelemans et al., 1999; Grefenstette, 1996)) which is easier to construct than a full-blown parser.

2.1.2 Verb Phrases as indexing terms

Semantically, a verb phrase can be seen as the description of a fact, event or process. It describes something dynamic, in contrast to the noun phrase which describes something static. The verb phrase in our interpretation comprises a verb group together with its complements. For the aboutness of the phrase, only the main verb is of importance, because the auxiliaries serve only to indicate time, modality, emotion. They will be elided during the transduction. The relations between the main verb and its complements however, including the subject, are essential for the aboutness of the phrase.

2.2 Phrase normalization

Phrases used as terms are very precise, much more precise than single words. The reason why it is hard to gain by using them, is the fact that they are so precise: the probability for a specific phrase to (re)occur in a document is much smaller than for a specific word. Thus, the term space for phrases is much more sparse than that of words. Using phrases, we gain Precision but we loose Recall.

In writing a text, people prefer to avoid literal repetition, they will go out of their way to choose another word, another turn of phrase, to use anaphora, synonymy, hypernymy, periphrasis, metaphor and hyperbole, in order to avoid the literal repetition of what they have written elsewhere. From a literary standpoint this is wonderful, but it gives complications in IR: we have to compensate for the human penchant for variety. Essentially, we have to conflate all semantically equivalent forms of a phrase, for instance by mapping them onto one same form.

Rather than using phrases taken from the text as terms, we shall reduce them to a normal form which expresses only the bare bones of their aboutness. We must eliminate all dispensable ornaments and undo all morphological, syntactic and semantic variation. In this way we strive to regain Recall while surrendering little Precision.

3 HM frames

In the context of document classification, we shall represent each document document by a bag of terms, where the terms are *Head/Modifier frames*, derived from the phrases in the document by a *transduction* process. Each phrase is transduced to one or more HM frames.

A Head/Modifier frame, HM frame or HMF is a pair denoted as

[head, modifier]

where both the head and the modifier consist of either one word extracted from the document or a sequence of such words and (recursively) frames.

The intuition behind this is that the modifier is joined to the head to make it more precise, i.e. to distinguish the frame from other frames with the same head, in particular to distinguish between different meanings of a polysemous head. Thus, there are many forms of engineering, but we may focus on [engineering, software].

It may be asked why we do not simply use *collocations* as terms, meaningful multi-word combinations, such as software engineering. Apart from the necessity for conflation of equivalent terms, for which HM frames give more scope, it should be pointed out that there is more information in the above HM frame than in the collocation: in expressing the relevance of a term to a class c we can use $P(c|head, modifier)$ in both cases, but $P(c|head, \overline{modifier})$ only for the HM frame – in this case the information that the document is about *another* form of engineering.

3.1 Unnesting

As produced by the transduction, frames may be nested, i.e. embedded in the head or modifier of another frame, but in the classification process only *atomic frames* will be used as terms, i.e. frames without embedded frames (possibly in combination with single words). These can be obtained by *unnesting* the original nested frames, as described in (Koster et al., 1999).

A frame may be nested in the head or the modifier of another frame, e.g.

```
[[ mower, lawn], large]
 [ operation, [consuming, time]]
```

Such a nested frame F may be unnested into a set of atomic frames S by repeatedly taking some embedded atomic frame from F while replacing it by its head, and adding that atomic frame to the set S , until F contains no further frames. Some (artificial) examples:

```
[a, b] ==> [a, b]
[a, [b, c]] ==> [b, c][a, b]
[[a, b], c] ==> [a, b][a, c]
```

In the unnesting process, the head of a frame serves as an abstraction for that frame: a lawn mower is a (some kind of) mower, a large mower is a mower, etc.

```
[[ mower, lawn], large] ==>
 [mower, lawn][mower, large]
```

We also allow the |-sign as an or-operator:

```
[a, b | c ] ==> [a, b][a, c]
[a | b, c ] ==> [a, c][b, c]
```

Finally, a frame with curly brackets instead of square brackets is used to delimit a frame, typically occurring nested in another one, which is to be unnested but with an empty abstraction, e.g.

```
{I, [see, man {man, [eat, chips | in garden]}}]
 ==> [I, see] [see, man] [man, eat]
      [eat, chips] [eat, in garden]
```

which is yielded from the sentence I saw this man eating chips in the garden.

4 Types of HM frames

Head/Modifier frames should not contain arbitrary combinations of words from the text, but they should represent some *linguistically meaningful relation* between components (words and collocations) extracted from the text.

4.1 Word types

It is to be expected that the words (and collocations) occurring in the HM frames should include the words that as keywords serve best to characterize the document; the frames only add precision. Traditionally, nouns are considered the most important keywords, and function words are eliminated by the stoplist. (Arampatzis et al., 2000) have investigated the relative contribution of words with different parts of speech to the accuracy in automatic document classification. It turned out that eliminating all terms except nouns, adjectives and verbs gives no loss (sometimes a small gain) in classification accuracy. These words appear to carry all of the aboutness to be found in single keywords.

We shall therefore restrict ourselves to four main types of words:

```
V: verb
N: noun
A: adjective
P: personal pronoun.
```

The pronouns are not by themselves very informative, but they have to be included because they often appear as placeholders. All other words like adverbs, auxiliaries, determiners etc will be elided.

4.2 Word relations

The set of all possible pairwise combinations of the four word types can be analysed as follows:

- verb in modifier position
a frame of the form [N:, V:] or [P:, V:] represents the *subject relation*. The type [A:, V:] cannot occur, since an adjective in that position would be promoted to a noun ([N:A:, V:]).
- verb in head position
a frame of the form [V:, N:] or [V:, P:] where the verb should be transitive represents the *object relation*. The type [V:, A:] can only occur for auxiliary verbs and therefore will not be generated. In a sentence like his nose turned red the verb does not provide any additional aboutness beyond [N:nose, A:red].

head	modifier				
	V	N	P	A	PP
V	–	object relation		–	yes
N, P	subject relation	predicative/attributive relation			yes
A	–	–	–	–	yes

Figure 1: Relations realized by HM frames

3. no verb

(a) noun as head

A noun can be modified by another noun [N:, N:], an adjective [N:, A:] or a (possessive) pronoun [N:, P:], all expressing the *attributive* relation (software engineering, the red car, my car). The same frames arise from the *predicative* relation (the car is a lemon, the car is red, the car is mine).

(b) adjective as head

Will not occur, since an adjective can only be modified by an adverb, which is elided.

(c) a pronoun in head position is treated like a noun.

4. verbs in both positions

Although sentences like I like to walk might suggest a frame [P:I, [V:like, V:walk]], this sentence has the same aboutness as I walk. Since we eliminate all modalities, verb/verb combinations will not be generated.

To complicate matters, we have also to represent indirect objects, preposition complements and adjuncts. For that purpose, we allow a verb, noun or adjective to have a modifier consisting of a preposition followed by a noun or pronoun which stands as the abstraction of an NP, for example

I give you a knife	[V:give, knife to P:you]
a cry for freedom	[N:cry, for N:freedom]
open to suspicion	[A:open, to N:suspicion]

Our HM frames therefore include the *index expressions* of (Bruza, 1993) as a subset.

Other relationships (negation, quantifiers, auxiliary verbs, adverbial modifiers) are not expressed in the transduction, even though they are recognized by the grammar, because there is some evidence that they are not important for the intended application in text classification (Arampatzis et al., 2000).

Notice that the *order* of the constituents in a frame is important for all the relations described above except the predicative/attributive one. We shall exploit the *polarity* in the frames to distinguish a frame [a, b] from a frame [b, a].

Expressed in the framework of HM pairs, the structure of a simple sentence will be

[subject, [verb, object | other complements]]

where some of the elements may be missing. By unnesting, this structure yields the atomic HM frames

[subject, verb], [verb, object] and [verb, other complement]

In a later stage, we shall investigate the effect of omitting the unnesting process and directly using the complicated tree-formed HM structures as terms.

4.3 Morphological normalization

All words or collocations occurring in the frames will be morphologically normalized by lemmatization. The goal is to map all different forms of the same lemma onto one representative form. The lemmatization process is aided by the word type supplied in the transduction.

As an example, the frame [N:man, A:V:sneezing] is obtained from both the attributive the sneezing man and the predicative the man is sneezing. It will be normalized to [man, sneeze]. The nested subject relation frame [N:man, [V:sneezed,]] obtained from this man sneezed has the same result, just as all its variants in time and modality.

All pronouns are mapped onto it, which can be seen as a placeholder element.

4.4 Syntactic normalization

As indicated above, all elements which are deemed not to contribute to the aboutness will be elided during transduction: articles, quantifiers, adverbs, connectors - much like applying a stop list. We shall have to determine experimentally which elements may be elided and what information should be expressed in the frames (e.g, time and modality). In our classification context, it is quite feasible to investigate for any specific feature whether its inclusion or exclusion would measurably influence the classification result. At a later time we will also investigate the use of non-atomic frames, by partly or wholly dispensing with the unnesting.

For each construct described by the grammar, its transduction is described wholly by the grammar (as part of the syntax rule for the

construct). Thus, the transductions of complicated constructs are expressed compositionally in terms of those of their components. In the process, elements are elided or re-ordered and additional symbols injected (like the [, , and] in order to express uniformly the four relations described above.

The syntactic normalizations implemented in this way include *em de-passivation*: the sentence
the train was driven by a clockwork engine
is, by unnesting and morphological normalization, turned into

```
[N:engine,N:clockwork] [N:engine,V:drive]
[V:drive,N:train]
```

4.5 An example

The sentence

```
Every PhD student gets a reduced price for the
ETAP conference
```

leads to the nested frame

```
[[N:student,N:PhD],[V:gets,[N:price,
{P:it,[V:reduced,N:price]}]
for [N:conference,N:ETAP]]]
```

Note the embedded sentence {P:it,[V:reduced,N:price]} rather than the simple [price,reduced], which is caused by the fact that *reduced* is the past participle of a transitive verb rather than a “neutral” adjective.

This complicated frame is then unnested to the set of atomic frames

```
[N:student,N:PhD] [N:student,V:gets]
[V:gets,N:price] [P:it,V:reduced]
[V:reduced,N:price] [N:price,
for N:conference] [N:conference,N:ETAP]
```

from which, after morphological normalization, the types are removed:

```
[student,PhD] [student,get] [get,price]
[it,reduce] [reduce,price] [price,for
conference] [conference,ETAP]
```

which should capture the aboutness of the sentence.

5 Some experimental results

In this section we give some rather preliminary experimental results concerning the distribution of HM frames in full text, obtained with the EP4IR parser/transducer (see 5.1). Of course these results are influenced by the accuracy of the grammar used. We shortly describe the EP4IR grammar used in the experiments, and introduce a technique for measuring the accuracy of a grammar in terms of the HM frames produced, and apply it to the EP4IR grammar.

5.1 The EP4IR grammar

The ‘English Phrases for IR’ (EP4IR) grammar and its associated lexicon were developed specifically for the extraction of phrases from English text and their transduction to Head/Modifier frames.

Since the grammar is not the main subject of this paper, we will only sketch its main properties. It gives a robust description of the structure of the Noun Phrase and the Verb Phrase, including their transduction to Head/Modifier frames. It has an extensive lexicon (309007 entries including collocations) uses various techniques to combat ambiguity (subcategorization, penalties).

5.2 Measuring grammar accuracy

In developing a grammar, there is a need for objective measures for the accuracy of the grammar, that allow us to do regression testing after each major modification to the grammar, and to assess the suitability of the parsers resulting from it for their intended purpose.

We can not use the ubiquitous Bracket Crossing (BC) measure (Marcus et al., 1994). To begin with, it has a large number of weakness and shortcomings (Carroll et al., 1999): it is not suited for partial parses, only useful for constituency based parsers, not fine grained enough for some specific syntactic phenomena, and there is no clear agreement on the granularity of bracketing. (Lin, 1995) has observed that in BC a mis-attachment can be punished more than once, so that a shallow parse with less syntactic information scores better than a “richer” analysis. Furthermore, the bracket-crossing approach needs an extensive syntactically analyzed corpus, which causes a chicken-and-egg problem, and it is very much oriented to parse trees, whereas we are interested in HM frames.

The HM frames generated from a text represent precisely the relations that we are interested in, and closely reflect the constituency or dependency relations in the text. If we can extract the right HM frames from a sentence, we can also derive the complete sentence structure. Therefore we propose to express the accuracy of the grammar/transducer in terms of the HM frames produced by it when applied to a reference corpus for which the HM frames are known.

This HMF-annotated reference corpus is based on a collection of sentences which is representative for the intended application domain and for the syntactic constructions occurring in it. This collection need not be very big (in comparison with a modern treebank) but it must be expected to generate enough HM frames to allow a measurement at the required granularity - say two thousand HM frames if we want 3 decimals of accuracy, a few hundred sentences.

The manual annotation of the corpus with the correct HM frames is tedious and errorprone, but

HMF	present	found	correct	recall	precision
VV	7	6	1	0.143	0.167
VN	178	139	89	0.500	0.640
VP	1	2	1	1.000	0.500
VA	10	15	3	0.300	0.200
VZ	87	63	28	0.322	0.444
NV	89	99	48	0.539	0.485
NN	257	212	171	0.665	0.807
NP	3	4	3	1.000	0.750
NA	220	192	150	0.682	0.781
NZ	111	131	53	0.477	0.405
PV	66	46	34	0.515	0.739
PN	1	1	0	0.000	0.000
PA	2	1	0	0.000	0.000
PZ	1	2	0	0.000	0.000
AN	2	0	0	0.000	0.000
AZ	2	1	1	0.500	1.000
total	1037	915	604	0.582	0.660

Figure 2: HM frames in mixed testset

it can be performed in interaction with the parser to be tested, presenting the sentences in larger or smaller fragments and verifying the results by inspection. By skillfully exploiting the compositional character of grammar, this can be done in an efficient and reliable way. Two persons may mark the same corpus and discuss the points of difference in order to guarantee the correctness of the reference corpus.

The measurement procedure is as follows:

1. generate a parser/transducer from the grammar and its lexicon
2. collect the reference corpus
3. manually derive the HM frames to be generated from it
4. let the parser generate HM frames from it
5. compare these, computing precision and recall in the usual way.

5.3 The accuracy of EP4IR

In order to get an indication of the accuracy (Precision and Recall) of EP4IR, we constructed a small reference corpus, consisting of 26 sentences from OHSUMED and 49 sentences from the EPO corpus, totalling 2245 words. By semi-automatic analysis, a total of 1037 HM frames was found to be present in the testset.

We analyzed the same testset automatically, the resulting distribution of HM frames is shown in fig. 2. As can be seen from the table, the overall Precision and Recall were 66.0% and 58.2%, giving an F1-value of 0.6188. The NN and NA combinations have the highest accuracy. At the moment, the grammar is not yet very accurate and there are still many inconsistencies in the transduction.

The breakdown of Precision and Recall per type of HM frame allows us to focus on areas of improvement. There are some erroneous combinations, like VV (also in the testset). PP attachment and the assignation of subject and object must be improved. But these figures are very initial, and need to be taken with some salt.

We also took a sample of newspaper prose (24210 words including punctuation marks; 147720 bytes) from the Wall Street Journal 1987 and extracted 10388 atomic HM frames from it, which showed roughly the same distribution of HM frame types.

The current version of EP4IR makes no use of probabilities, and therefore in case of lexical or syntactic ambiguity is not good at finding the most likely analysis. Furthermore, it is weak in its treatment of coordination and some special constructs.

5.4 Invitation

The EP4IR grammar is presently being brought under the GPL, and we invite the reader to use it. It gives you a head start in researching and developing NLP-based applications and IR systems. And in case you are not satisfied about its accuracy, you are kindly invited to contribute to its improvement.

6 Concluding remarks

We have introduced HM frames, described their use for document representation and introduced a linguistically motivated taxonomy of HM frames allowing to capture the aboutness (as opposed to the constituency structure) of both the verb phrase and the noun phrase while rigorously eliminating non-informative elements.

We have introduced an accuracy measure for grammars transducing to HM frames and mea-

sured the Precision and Recall achieved by the EP4IR parser/transducer on a small test corpus. There is obviously room for improvement, and we are working on it.

It is an open question whether HM frames (and LMT's in general) should be used as terms on their own, or whether additionally some or all of the words of the document should be used as terms.

This question can be resolved experimentally now that the HMF transduction is operational. Using our classification engine (Koster et al., 1999), the accuracy and performance achieved with any specific choice of document representation can be measured, comparing different classification algorithms and cross-validating on many different corpora.

Other goals for our future research are:

- to determine experimentally which elements of a sentence contribute most to the aboutness of a document
- to investigate theoretically how to make optimal use of structured terms such as our HM frames in text classification
- to investigate the possibility to (partially) dispense with unnesting, using arbitrarily complicated LMT's
- to investigate anaphora resolution with HM frames, selective clustering of terms and fuzzy semantic matching.

The EP4IR grammar of English used in the above experiments, which was developed for investigations into the effective use of phrasal document representations in Information Retrieval applications like document classification, filtering and routing, is now available in the public domain, together with its lexicon and the AGFL parser generator system.

References

- A. Arampatzis, Th.P. van der Weide, C.H.A. Koster, P. van Bommel (2000), An Evaluation of Linguistically-motivated Indexing Schemes. Proceedings BCS-IRSG 2000 Colloquium on IR Research, Cambridge, England.
- P. Bruza and T.W.C. Huibers, (1994), Investigating Aboutness Axioms using Information Fields. Proceedings SIGIR 94, pp. 112-121.
- P. Bruza and T.W.C. Huibers, (1996), A Study of Aboutness in Information Retrieval. *Artificial Intelligence Review*, 10, p 1-27.
- P.D. Bruza (1993), *Stratified Information Disclosure: A Synthesis between Information Retrieval and Hypermedia*, PhD thesis, University of Nijmegen.
- J. Carroll, M. Guido and E. Briscoe (1999) Corpus Annotation for Parser Evaluation. *Proceedings of the EACL workshop on Linguistically Interpreted Corpora (LINC)*, 1999.
- W. Daelemans, S. Buchholz and J. Veenstra (1999), Memory-based shallow parsing, proceedings CoNLL, Bergen, Norway.
- D.A. Evans, R.G. Lefferts, G. Grefenstette, S.H. Handerson, W.R. Hersch and A.A. Archbold (1993), CLARIT TREC design, experiments and results. TREC-1 proceedings, pp. 251-286.
- J.L. Fagan (1988), *Experiments in automatic phrase indexing for document retrieval: a comparison of syntactic and non-syntactic methods*, PhD Thesis, Cornell University.
- G. Grefenstette (1996), Light parsing as finite state filtering. Workshop on Extended finite state models of language, Budapest, ECAL'96.
- C.H.A. Koster, Affix Grammars for Natural Languages. In: H. Alblas and B. Melichar (Eds.), *Attribute Grammars, applications and systems*. SLNCS 545, Heidelberg, 1991, p. 469-484.
- C.H.A. Koster, C. Derksen, D. van de Ende and J. Potjer, Normalization and matching in the DORO system. Proceedings BCS-IRSG 1999 colloquium, Glasgow University.
- D. D. Lewis (1992), *Representation and Learning in Information Retrieval*. PhD thesis, Department of Computer Science; Univ. of Massachusetts; Amherst, MA 01003.
- D. Lin (1995), A dependency-based method for evaluating broad-coverage parsers. *Proceedings IJCAI-95*, pp. 1420-1425.
- M. Marcus, B. Santorini and M. Marcinkiewicz (1994), Building a Large Annotated Corpus of English: The Penn Treebank. In: *Computational Linguistics*, 19(2):313-330, 1994.
- C.J. van Rijsbergen (1979), *Information Retrieval*, Butterworth. London, second edition.
- A.F. Smeaton (1997), Using NLP and NLP resources for Information Retrieval Tasks. In: T. Strzalkowski (Ed.), *Natural Language Information Retrieval*, Kluwer Academic Publishers.
- K. Sparck Jones (1998), Information retrieval: how far will *really* simple methods take you? In: Proceedings TWTL 14, Twente University, the Netherlands, pp. 71-78.
- K. Sparck Jones (1999), The role of NLP in Text Retrieval. In: (Strzalkowski, 1999) pp. 1-24.
- T. Strzalkowski (1995), Natural Language Information Retrieval, *Information Processing and Management*, 31 (3), pp. 397-417.
- T. Strzalkowski, editor (1999), *Natural Language Information Retrieval*, Kluwer Academic Publishers, ISBN 0-7923-5685-3.
- T. Winograd (1983), *Language as a Cognitive Process: Volume I: Syntax*, Reading MA, Addison-Wesley, 650 pp.