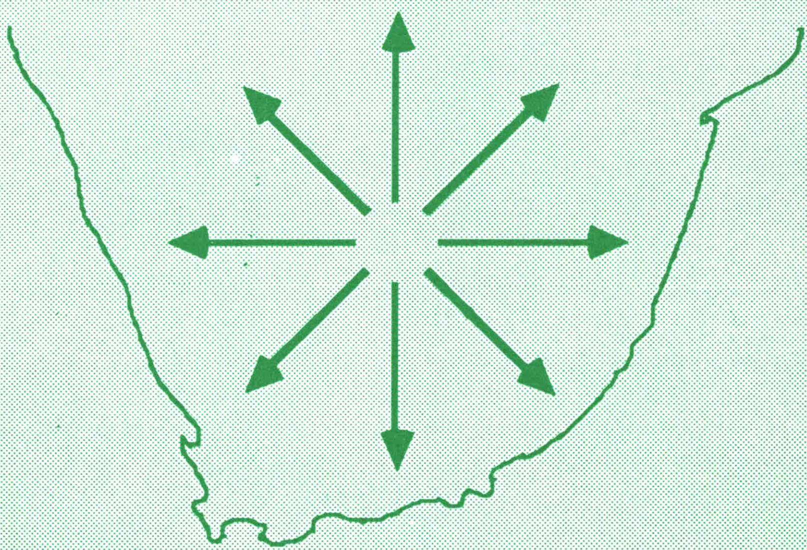
South African Computer Journal
Number 7
July 1992

Suid-Afrikaanse Rekenaartydskrif
Nommer 7
Julie 1992

SPECIAL ISSUE



7th SA COMPUTER RESEARCH SYMPOSIUM

The South African Computer Journal

An official publication of the South African Computer Society and the South African Institute of Computer Scientists

Die Suid-Afrikaanse Rekenaartydskrif

'n Amptelike publikasie van die Suid-Afrikaanse Rekenaarvereniging en die Suid-Afrikaanse Instituut vir Rekenaarwetenskaplikes

Editor
Professor Derrick G Kourie
Department of Computer Science
University of Pretoria
Hatfield 0083
Email: dkourie@dos-lan.cs.up.ac.za

Subeditor: Information Systems
Prof John Schochot
University of the Witwatersrand
Private Bag 3
WITS 2050
Email: 035ebrs@witsvma.wits.ac.za

Production Editor
Prof G de V Smit
Department of Computer Science
University of Cape Town
Rondebosch 7700
Email: gds@cs.uct.ac.za

Editorial Board

Subscriptions

	Annual	Single copy
Southern Africa:	R45,00	R15,00
Elsewhere	$45,00	$15,00

to be sent to:
Computer Society of South Africa
Box 1714 Halfway House 1685

Computer Science Department, University of Pretoria, Pretoria 0002
Phone: (int)+(012)+420-2504 Fax: (int)+(012)+43-6454 email: dkourie@rkw-risc.cs.up.ac.za

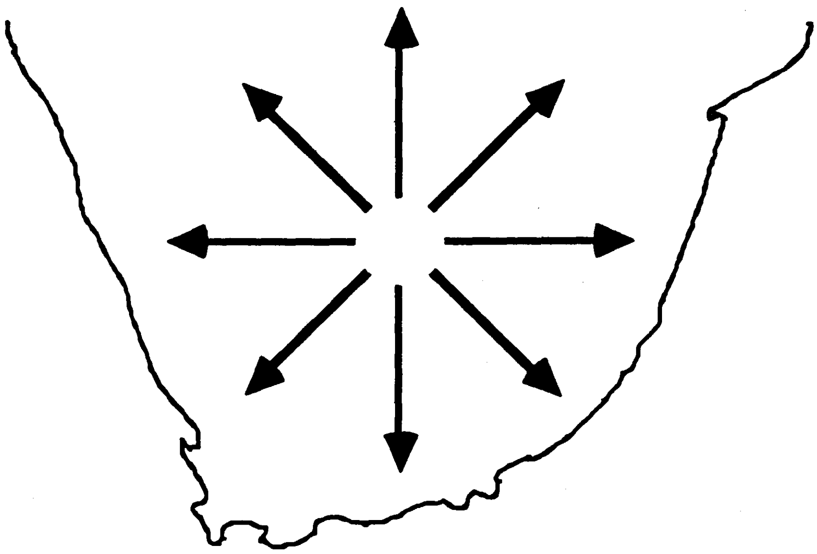
7th Southern African Computer Research Symposium

Karos Indaba Hotel, Johannesburg

1 July 1992

PROCEEDINGS

Guest Editor: Judy M Bishop



PERSETEL

Organised by the SA Institute of Computer Scientists
in association with the Computer Society of SA

Sponsored by Persetel and the FRD

SPECIAL ISSUE - 7th SA COMPUTER SYMPOSIUM

PREFACE

When the first SA Computer Symposium was held at the CSIR in the early eighties, it was unique. There was no other forum at the time for the presentation of research in computer science. In the intervening decade, conferences, symposia and workshops have sprung up in response to demand, and now there are several successful ventures, some into their third or fourth iteration. Each of these addresses a specific topic - for example, hypermedia, expert systems, parallel processing or formal aspects of computing - and attracts a specialised audience, well versed in the subject and eager to learn more. For the main part, the proceedings are informal, and certainly not archival.

SACRS, though, is still unique, in that it deliberately covers a broad spectrum of research in computing, and in addition, seeks to provide a lasting record of the proceedings. To achieve the second aim, we negotiated with the SA Institute of Computer Scientists for the proceedings to form a special issue of the SA Computer Journal, and the copy you have in front of you is the result. The collaboration between the symposium committee and the journal's editorial board placed high standards on the refereeing and final presentation of the papers, to the symposium's benefit, while we were still able to maintain a fresh, audience-oriented approach to the selection of papers.

This is SACJ's first such special issue, and the largest issue (at 145 pages) to date. We hope that it is only the beginning of future such collaborations.

In all 29 papers were received, all were refereed twice, and 19 were chosen for presentation by the programme committee. All the papers were thoroughly revised by the authors on the basis of the referee's comments, and the committee's suggestions aimed at making the material more accessible to a broadly-based audience. Papers had to be new, and not to have been presented elsewhere, a requirement that is still unusual within the SA conference round.

A third goal of SACRS has been to invite keynote speakers, usually from overseas. This year, we are fortunate to present Dr Vinton Cerf, the father of the Internet and a world-renown expert on computer networks. Although his paper is not available for this special issue, it will appear later in SACJ. Through the good offices of Professor Chris Brink of UCT, we also have three other speakers from Germany, Canada and the US adding interest to the event, and two of their papers appear in this issue.

The programme committee originally devised a theme for the symposium - "Computing in the New South Africa". We received several queries as to the meaning of this theme, but unfortunately few papers that addressed it directly. One prospective author went as far as to enquire whether computer research would survive in the new South Africa. Another felt that his work was definitely not in the theme, as it was genuine, old world, basic, theoretical science! Neverthless, there are two papers that consider one of South Africa's key issues, that of language. Others look at the success we have achieved in applying technology to mining, and the future of low-cost operating systems. In all, the mix of papers represents a balance between the theoretical and the practical, the past and the future, all firmly based in the computing of the present.

Organising the symposium has involved the hard work of several people, and I would like to thank in particular

- Derrick Kourie, my co-organiser, and the editor of SACJ for his invaluable advice and hard work throughout the planning and implementation stages;
- Riël Smit, the production editor, for attaining such a high standard in such a short time for so many papers;
- Gerrit Prinsloo and the staff at the CSSA for their efficeint and quite delightfully unfussy organisation;
- Persetel for their very generous sponsorship of R25000, and Tim Schumann for taking a genuine interest in our events;
- the Foundation for Research Development for sponsoring Vint Cerf's visit;
- and finally the Department of Computer Science of the University of Pretoria for providing the ideal working conditions for undertaking ventures of this kind, and especially Roelf van den Heever for his unfailing encouragement and support.

Judy M Bishop
Organising Chairman, SACRS 1992
Guest Editor, SACJ Special Issue

Referees

The journal draws on a wide range of referees. The following were involved in the refereeing of the papers selected for this special issue. Their role in certifying the papers and their contribution to enhancing the quality of papers is sincerely appreciated.

John Barrow	UNISA
Ronnie Becker	University of Cape Town
Danie Behr	University of Pretoria
Sonia Berman	University of Cape Town
Liesbeth Botha	University of Pretoria
Theo Bothma	University of Pretoria
Chris Brink	University of Cape Town
Peter Clayton	Rhodes University
Ian Cloete	Stellenbosch University
Antony Cooper	CSIR
Elise Ehlers	Rand Afrikaanse University
Quintin Gee	DALSIG
Andy Gravell	University of Southampton
Wendy Hall	University of Southampton
David Harel	The Weizmann Institute of Science
Scott Hazelhurst	University of the Witwatersrand
Derrick Kourie	University of Pretoria
Willem Labuschagne	UNISA
Doug Laing	ISM
Dave Levy	University of Natal
Graham McLeod	University of Cape Town
Hans Messerschmidt	University of the Orange Free State
Deon Oosthuizen	University of Pretoria
Ben Oosthuysen	University of Pretoria
Neil Pendock	University of the Witwatersrand
D Petkov	University of Natal
Martin Rennhackkamp	Stellenbosch University
Cees Roon	University of Pretoria
Jan Roos	University of Pretoria
John Shochot	University of the Witwatersrand
Morris Sloman	Imperial College, London
Riel Smit	University of Cape Town
Pat Terry	Rhodes University
Walter van den Heever	University of Pretoria
Lynn van der Vegte	University of Pretoria
Herman Venter	University of Fort Hare
Herna Viktor	Stellenbosch University

Automatically Linking Words and Concepts in an Afrikaans Dictionary

P Z Theron I Cloete
Department of Computer Science, University of Stellenbosch, Stellenbosch 7600, South Africa
Internet: ian@cs.sun.ac.za

Abstract

The problem of automatic acquisition of lexical-semantic relations for Afrikaans nouns from a dictionary is addressed. The acquisition process is improved over previous approaches by implementing both typographic constraints in patterns, and by automating the acquisition of syntactic patterns. Results for the noun taxonomy relation IS-A and semantic feature IS-HUMAN show that ambiguity can successfully be resolved with the patterns obtained.
Keywords: *computational lexicography, lexical-semantic relations, natural language acquisition.*
Computing Review Categories: *I.2.6, I.2.7*

Received March 1992, Accepted April 1992

1 Introduction

Natural language (NL) processing systems can play an important role in the new South Africa in various applications such as education, interfacing with computers, information retrieval, question answering, dissemination of information and machine translation. Any NL system providing broad language coverage requires a detailed lexicon containing explicit information about its target language. Such a lexicon needs to be much larger than those currently available. Today's NL systems are restricted to a particular sublanguage because the lexicon has to be handcoded — a very time-consuming process. The only viable alternative sources of lexical information are written texts such as dictionaries, which contain a wealth of explicit and implicit information. A commercial dictionary, carefully constructed by lexicographers, is a product of several man years of language expertise and serves as an excellent basis from which to construct a lexicon. The research described in this paper is part of an effort to automate the construction of an Afrikaans lexicon from machine-readable dictionaries. We chose Afrikaans because it is a local language for which a NL system would be useful, a machine-readable dictionary [3] was available and, to our knowledge, no similar project has been done for Afrikaans [7].

In this paper we consider the extraction of implicit semantic links between words in the noun entries of the dictionary. An example of a very common semantic link is the taxonomy relation (IS-A) (e.g. *lamb* IS-A *sheep*). These implicit links must be encoded in a structured form by deriving explicit relations linking concepts to related words. Such relationships may, for instance, serve to disambiguate the possible interpretations of a word in a particular context.

To this end, we present an automated methodology based on *defining formulae*, and also make use in a novel way of syntactic and typographic patterns present in the meaning descriptions of the dictionary. To illustrate the methodology, preliminary results are given for the three relations IS-A (taxonomy), PART-OF (part-whole) and IS-HUMAN (also called a *semantic feature*). Of these three, the extraction of IS-A and IS-HUMAN has been the most successful, while for PART-OF further work has to be done to obtain acceptable accuracy. We expect that our methodology and the link types extracted would extend to other languages as well, since many semantic relations encode knowledge about the world, which is language invariant. Furthermore, if *source language – target language* synonyms are available, many instances of the semantic relations would directly transpose to other (target) languages.

We start with a brief explanation of lexical-semantic relations and defining formulae, as well as their use in previous work. Then our methodology for the extraction of relations is presented. We conclude with the results obtained for the three relations. A short Afrikaans/English glossary is provided in an appendix.

2 Lexical-Semantic Relations

A lexical-semantic relation (LSR) is any specialized link existing between lexical items and/or concepts [8]. An example of a set of links is the association of a word with its different meanings. These links form the basis for the construction of dictionary entries. LSRs are generally present as morphological, syntactic and semantic relationships, but other relationships such as propositional attitude relations (e.g. implicatures [8]) also occur. Researchers have identified more than a hundred different LSRs through theoretical analysis and the analysis of dictionaries. The interested reader is refered to [4] for a comprehensive discussion of LSRs and to [8] for an abbreviated LSR list, as well as an example of an information retrieval application.

LSRs are present in the dictionary as *explicit* and *implicit* information. An example of an explicit LSR is the synonymy relation. All the other major semantic LSRs, inluding the three we consider in this paper, are defined implicitly. The problem is to make them explicit.

The utility of these LSRs in NL systems is not discussed here, and a quick example will suffice to illustrate their use. The following sentence is normally considered to be semantically incorrect since the verb *"drives"* expects a human subject.

*The tree drives the car.

The fact that *"tree"* does not satisfy the IS-HUMAN relation must be determined in an NL system from an explicitly coded IS-HUMAN LSR.

An example of the way in which these LSRs are present in a dictionary entry, is given below:

bakker (-s) *s.nw. 1. persoon wat brood, koek ens. bak. 2. eienaar of bestuurder van 'n bakkery of bakkerswinkel.*

In this entry, the defined word *"bakker"* has the syntactic category noun (indicated by *"s.nw."*) and two associated meaning descriptions (MDs). Three LSRs occur in the example:

1. The first MD starts with the word *"persoon"*, which allows the association of the feature IS-HUMAN with the defined word.
2. The second MD contains two implicit IS-A relations:
   (a) *bakker* IS-A *eienaar*
   (b) *bakker* IS-A *bestuurder*

LSRs which have the defined word as the first argument and the *semantic head* of the first phrase in the MD as the second argument, are called *primary relations*. The *semantic head* of a phrase can be described as the word carrying most of the meaning of the phrase.

Further relations between the defined word and other words in the MD are named *secondary*, and so forth. Note that both IS-A relations in the above example are primary relations because the second MD has two arguments separated by a disjunction *"of"*.

IS-A LSRs are very common in the dictionary because there is a close correspondence between them and *genus-differentia* type MDs which are used on a large scale in the dictionary [5]. A genus-differentia MD defines a word by

1. specifying a semantic class (the genus term) to which the concept represented by the word belongs, and by
2. providing a number of features (the differentia) which distinguish the defined concept from other concepts in the same class.

For example, in the second MD of *"bakker"* given above, the genus term is either *"eienaar"* or *"bestuurder"* and the concept *"bakker"* is distinguished from all other *"eienaars"* and *"bestuurders"* by qualifying it with the phrases *"van 'n bakkery of bakkerswinkel"*. Note that for this type of MD the primary relations associate the defined word with the class(es) to which it belongs while the defined word and its differentia are linked by secondary and further relations. In this paper we are concerned only with the primary relations.

3 Defining Formulae

Thus, in order to extract the implicit LSRs from a dictionary entry, it is necessary to analyze the meaning descriptions. One approach to this problem, is to make use of *defining formulae* (DFs). Defining formulae are certain phrases which occur frequently in the meaning descriptions and are good clues to LSRs. It can be a single word like *"soort"* or a phrase such as *"gedeelte van 'n"*. Other researchers have already demonstrated the utility of DFs in the identification of possible LSRs, e.g. [1] and [6]. In the example dictionary entry given above, the word *"persoon"* is one of the identified DFs that indicates the feature IS-HUMAN. An example of a DF indicating the PART-OF relation is *"gedeelte van 'n"*. The typical approach to identify DFs is to use a program which finds recurring phrases of a high frequency in MDs [1].

Because the relationship between DFs and LSRs is generally many to many and not one to one, the association of which DF indicates which LSR must be further constrained by considering the valid *arguments* of an LSR. In the example dictionary entry given above, the first and second arguments of the first IS-A relation are *"bakker"* and *"eienaar"* respectively. The first argument is a noun since only noun entries are considered. Thus, a syntactic constraint on the second argument of an IS-A relation, namely that it must be a noun, would further restrict the matching process and may also serve to identify the argument.

These arguments appear in the surrounding contexts of the DFs, i.e. the words preceding and following the DF. Other researchers obtained a listing of these contexts with the aid of a key word in context (KWIC) program. These listings then had to be manually perused to determine the arguments of an LSR. Another approach is to construct a grammar for the MDs and use parsing to extract the arguments. The result of this process is a list of $argument_1$-relation-$argument_2$ triples like the following:

ewenaar	IS-A	*middellyn*
flank	PART-OF	*dier*
diplomaat	HAS-FEATURE	*IS-HUMAN*

These triples are the building blocks of a *relational lexicon* [8].

4 Methodology

To perform this research we wrote a parser in PROLOG for the machine-readable text provided by the publishers. Typesetting information, such as font changes from bold to cursive, made it possible to distinguish dictionary entries and the fields of each entry. This information was stored as normalized relations in a Relational Database. The database includes for each defined word its

- alternative spellings (if any),
- syntactic category (part of speech),
- morphemes for constructing derivatives (e.g. the plural form of a noun),
- meaning descriptions and
- optional examples of the usage of the word.

The three LSRs for Afrikaans nouns were determined in several stages of processing. All MDs were augmented by adding two special tokens indicating the beginning (BOS) and end (EOS) of the MD respectively. A program, REP,

Table 1. A sample of the most frequently occurring candidate DFs.

Freq.	Formula	Freq.	Formula
8497	van	511	BOS soort
7314	'n	508	ander
6563	die	498	gebruik
4159	wat	482	, veral
3797	of	477	iets
2831	in	457	is .
2700	en	456	is . EOS
2502	van 'n	439	BOS iem.
2477	met	420	een
1710	van die	397	iem. wat
1547	word	367	BOS iem. wat
1496	te	361	groot
1471	vir	355	mens
1341	om	348	wat die
1127	op	342	na
1117	deur	337	bv .
921	word .	324	wat 'n
917	word . EOS	313	deel
915	is	311	tussen
858	aan	310	waarin
841	as	309	klein
825	in die	305	op 'n
715	met 'n	302	BOS persoon
712	soort	301	ens . EOS
697	by	296	oor
678	iem.	295	kan
654	in 'n	288	, ens . EOS
620	ens .	288	se
590	, ens	288	sy
587	, ens .	287	vir die
564	veral	285	wat in
541	uit	276	twee
532	persoon	275	deur 'n
517	nie	269	aan die

was written to automatically select likely candidates for DFs. REP receives the MDs as input and can then produce a list of the highest frequency recurring phrases of all desired lengths. We restricted the lengths of the DFs to those consisting of 1 to 7 tokens. These phrases are ranked according to their frequency of occurrence, and the highest frequency phrases are considered as candidates for defining formulae. A small sample of the DFs found is given in table 1.

Inspection of the 100 highest frequency DFs indicated that

1. DFs are still unacceptably ambiguous and need further constraining.
2. DFs appearing in the leftmost position of an MD or directly following a semicolon, are much less ambiguous then others which are preceded by at least one word.
3. Very short MDs occur which indicate the IS-A relation accurately, but contain no DF.
4. Extraction of the second argument requires a form of syntactic analysis.

To address these problems, two approaches were followed. The first approach used only the typographic information present in the MDs. Two wildcards were implemented: "WORD" which matches any natural language word and "DF" which matches any DF. These wildcards and the BOS and EOS tokens allowed the recognition of *typographic patterns* providing information on the relative positions of words to these two special markers, punctuation marks and the previously identified DFs. Some of these patterns indicate LSRs, even though no distinction is made between words or word classes. An example pattern indicating the IS-A LSR unambiguously is

BOS WORD . EOS

The second approach added syntactic information to the typographic information. A form of grammar learning was implemented, using only syntactic categories of words (and their computable derivatives) from the dictionary database to form syntactic patterns of words surrounding a DF. Other authors [1, 2] also used syntax to constrain and extract arguments of a DF, but (a) their grammar was handcoded for the *entire* MD, and (b) the construction process of the grammar was incremental — when the parser failed the grammar was updated. We observed that the syntax of the words surrounding a DF is limited in variation and that adequate syntactic patterns could be obtained by replacing a word with a derived token indicating all its category variations found in the dictionary database. This process produced patterns containing DFs as well as syntax constraints, and the actual arguments were thus easily obtained by doing a second processing pass on the MDs without requiring a grammar to be constructed. An example of such a pattern to be matched against the MDs is:

BOS NOUN DF

This way of acquiring syntax constraints is not without its problems, because

1. some words are not yet tagged with a syntactic category — there are 14653 words (55.6% of total) with unknown category which appear 41307 times in the MDs of all the nouns,
2. some words are tagged with more than one syntactic category since a word is considered without its surrounding context, and
3. some syntactic categories are obviously incorrect also due to the lack of context.

The fact that more than half of the words appearing in the MDs are not yet tagged with a syntactic category is mainly due to limited morphological analysis to determine the catagory of a derived word. Morphological analysis does not yet cater for all word variations. Currently simple derivatives of defined words are computed e.g. past participles of verbs or noun superlatives. A minor secondary reason is that a small number of dictionary entries could not be parsed due to inconsistent dictionary format. These problems are currently being addressed.

5 Results

Results for each of the three relations considered are presented in the following subsections. A total of 24295 noun MDs were successfully extracted from the dictionary. Percentages are given relative to this number. If the number of pattern matches for any given pattern against these MDs is large, the correctness of LSR extraction was estimated by drawing a small random sample of these matched patterns and checking them by hand. This is indicated in the tables by the column "Correct/Sample".

Table 2. The 10 most frequently occurring typographic patterns extracting IS-A 2nd arguments, with the positions of the arguments underlined.

Freq.	Typographic Pattern	Correct/ Sample
7291	BOS WORD DF	183/200
3166	BOS WORD . EOS	200/200
1806	; WORD . EOS	200/200
1310	BOS WORD , WORD . EOS	197/200
337	BOS WORD ; WORD	100/100
285	; WORD , WORD . EOS	88/100
284	BOS WORD , WORD , WORD , WORD . EOS	279/284
130	BOS WORD ; WORD . EOS	130/130
107	; WORD ;	106/107
96	BOS soort WORD . EOS	96/96

Table 3. The 10 most frequently occurring syntactic patterns extracting IS-A 2nd arguments, with the positions of the arguments underlined.

Freq.	Syntactic Pattern	Correct/ Sample
3304	BOS NOUN DF	93/100
1651	BOS NOUN . EOS	100/100
1002	; NOUN . EOS	100/100
683	BOS NOUN-or-VERB DF	94/100
269	BOS NOUN , NOUN .	269/269
199	BOS NOUN ;	197/199
125	; NOUN DF	120/125
94	BOS NOUN-or-VERB . EOS	94/94
92	BOS NOUN , NOUN DF	87/92
82	; NOUN , NOUN . EOS	79/82

Table 4. The 7 most frequently occurring DFs indicating the feature IS-HUMAN.

Freq.	Formula
439	BOS iem.
367	BOS iem. wat
302	BOS persoon
225	BOS persoon wat
41	BOS iem. wat 'n
31	BOS iem. met
31	BOS vrou

IS-A

IS-A LSRs were found using two sets of patterns, typographic and syntactic. Each of these sets of patterns are dealt with separately below. Both the determined lists of typographic and syntactic patterns consisted mainly of simple patterns indicating a primary IS-A relation. Table 2 contains 10 typographic patterns yielding IS-A LSRs. These patterns were selected conservatively from the 100 most frequently occurring patterns such that the resulting IS-A relation is correct with a high degree of certainty. These simple patterns provided a surprisingly large number of $argument_1$-IS-A-$argument_2$ triples. The typographic patterns match 14812 (61%) of the total noun MDs allowing the extraction of 14942 triples.

An inspection of 1617 triples matched by the patterns showed that 97,6% were correct. A breakdown by pattern of this number is given in table 2. Incorrect matches (indicated by *) were primarily due to the following:

- A semantic feature is selected as the 2nd argument, e.g.:

  dignitaris *IS-A *persoon*

  This can be avoided by marking these features in the MDs. It must still be determined whether such a selection is necessarily incorrect.
- Part of a DF is selected as the 2nd argument, e.g.:

  voet *IS-A *gedeelte*

  This can be avoided by marking the DFs in the MDs.
- Abbreviations which were not tokenized by the initial Prolog parsing of the dictionary tape, were selected by patterns containing *"WORD ."*.
- A few 1st and 2nd arguments were not complete words and morphological analysis is necessary to reconstruct them.
- The lack of syntax constraints caused 9 incorrect selections of adjectives in the sample of 200 triples matched by *"BOS WORD DF"*.

Table 3 shows the 10 most frequently occurring syntactic patterns. They match 7501 (30,8%) of the total noun MDs and provide a total of 7944 $argument_1$-IS-A-$argument_2$ triples. Out of the sample of 1261 triples in table 3, 1233 (97.8%) were found to be correct. These numbers can easily be increased by considering the less prominent patterns as well. The errors found in these triples are of the same kind as those found with typographic patterns, except for those errors due to a lack of syntax constraints.

For these relatively simple patterns, the typographic patterns allowed the retrieval of almost double the amount of triples compared to those made available by the syntactic patterns at a comparable accuracy. The reason for this is twofold:

1. The lexicographers made use of typographic regularity to convey information due to the concise nature of the dictionary, and
2. currently less than half of the words appearing in the MDs can be mapped to their possible syntactic categories as explained before.

IS-HUMAN

Table 4 shows the DFs which were obvious indicators of the IS-HUMAN feature. They were hand selected from the 100 most frequently occurring DFs appearing at the start of an MD. (Note that a feature does not require the determination of a second argument for the LSR.)

The DF *"BOS iem."* subsumes the DFs *"BOS iem. wat"*, *"BOS iem. wat 'n"* and *"BOS iem. met"*. Also, *"BOS persoon"* subsumes *"BOS persoon wat"*. Thus it is only necessary to verify the use of the DFs *"BOS iem."*, *"BOS persoon"* and *"BOS vrou"*, as indicators for the IS-HUMAN feature. Of the 439 selections made with *"BOS iem."* only the following four were incorrect:

- *handskrif : iem. se besondere manier van skryf.*
- *handtekening : iem. se voorletters en van deur homself geskryf.*
- *agtergrond : iem. se familie, jeugomstandighede, ens.*
- *boedel : iem. se hele vermoë met die daarbybehorende regte en verpligtings.*

All 302 selections using the pattern *"BOS persoon"* and the 31 selections using *"BOS vrou"* were found to be correct.

The automatic extraction of a number of other features seems trivial. Many of the DFs appearing at the start of an MD are likely candidates for features (see table 5). For example, the features IS-INSTRUMENT and IS-PLANT

Table 5. A sample of the most frequently occurring DFs appearing at the start of MDs.

Freq.	Formula	Freq.	Formula
511	BOS soort	60	BOS handeling van
439	BOS iem.	60	BOS plek waar
367	BOS iem. wat	59	BOS kort
302	BOS persoon	59	BOS studie
225	BOS persoon wat	59	BOS wat
218	BOS die	58	BOS groep
180	BOS klein	58	BOS lang
179	BOS groot	51	BOS enigeen van die
131	BOS enigeen	50	BOS dun
130	BOS enigeen van	50	BOS stuk
123	BOS iets	50	BOS wetenskap
114	BOS deel	48	BOS studie van
114	BOS toestel	46	BOS lid
109	BOS een	45	BOS lid van
104	BOS deel van	45	BOS sterk
100	BOS plek	44	BOS ligte
93	BOS een van	44	BOS toestand van
86	BOS leer	44	BOS fyn
81	BOS iets wat	43	BOS geld
80	BOS 'n	42	BOS harde
72	BOS handeling	41	BOS hoeveelheid
70	BOS toestand	41	BOS leer van die
69	BOS deel van 'n	41	BOS versameling
69	BOS plant	41	BOS iem. wat 'n
64	BOS baie	39	BOS plat
61	BOS leer van	37	BOS boonste

are indicated by DFs such as *"BOS toestel"* (occurring 114 times) and *"BOS plant"* (occurring 69 times).

PART-OF

In contrast to the IS-A relation which appears very often as a primary LSR in the MDs, the PART-OF relation is mostly present as a secondary or *internal* relation. By an *internal relation* we mean an LSR between two words, both of which occur in the MD. This complicates the decision of whether syntactic constraints are sufficient for automatic argument retrieval.

One of the DFs thought to indicate the PART-OF LSR, *"van"*, turned out to be very ambiguous. This is not surprising since it is the DF which occurred the most (8497 times — see table 1). Two related DFs *"van 'n"* and *"van die"* occurred 2502 and 1710 times respectively. Obviously there is direct correlation between the diversity of use of a DF and its degree of ambiguity; frequent use in diverse contexts indicates high ambiguity. To illustrate the ambiguity present in DFs for the PART-OF LSR, consider the pattern

BOS NOUN *van 'n* NOUN

Three MDs matching the pattern appear below.

1. *draketand : tand* van 'n *draak*
2. *evangelis : skrywer* van 'n *evangelie*
3. *diktatuur : regering* van 'n *diktator*

The first MD defines three relations, two of which are correct PART-OF instances:

- Primary Relation: *draketand* IS-A *tand*
- Secondary Relation: *draketand* PART-OF *draak*
- Internal Relation: *tand* PART-OF *draak*

The second MD does not contain the PART-OF LSR at all.

The third contains the PART-OF relation, but with the arguments appearing in reversed positions in the dictionary entry (i.e. the second argument precedes the first argument in the entry):

- Primary Relation: *diktatuur* IS-A *regering*
- Secondary Relation: *diktator* PART-OF *diktatuur*
- Internal Relation: *diktator* PART-OF *regering*

There are, however, some DFs appearing at the start of a meaning description which are good candidates for the automatic retrieval of the PART-OF LSR. These include the DF *"BOS deel"* and its subsumed DFs *"BOS deel van"* and *"BOS deel van 'n"* for which results are not yet finalized. We hope to report further results on the extraction of the PART-OF relation in the near future.

6 Further Research

The task at hand is to refine the extraction of internal LSRs, particularly the PART-OF relation, and to determine further semantic features (e.g. IS-INSTRUMENT).

Our methodology is also intended to be applicable to the extraction of LSRs present in the MDs of the other major parts of speech — verbs, adjectives and adverbs. One useful LSR which might be extracted from the verb MDs is a selectional marker indicating that the verb expects a human subject. Work on this will start in the near feature.

7 Conclusion

The results obtained verify the viability of automating the acquisition of lexical-semantic relations. It is, to our knowledge, the first such project launched for Afrikaans. We introduced a new approach in which typographic and syntactic patterns are acquired directly from the meaning descriptions, instead of being handcoded. The use of these patterns significantly constrained the matching of possible LSRs, making their automatic retrieval possible.

References

1. T Ahlswede and M Evens. 'Generating a Relational Lexicon from a Machine-Readable Dictionary'. *International Journal of Lexicography*, 1(3):214–237, (1989).
2. H Alshawi. 'Analysing the dictionary definitions'. In B Boguraev and T Briscoe, eds., *Computational Lexicography for Natural Language Processing*, chapter 7, pp. 153–170. Longman, London, (1989).
3. M de Villiers, J Smuts, L C Eksteen, R H Gouws. *Nasionale Woordeboek*. Nasou Beperk, Kaapstad, 1985.
4. M W Evens, ed. *Relational Models of the Lexicon*. Cambridge University Press, Cambridge, 1988.
5. R H Gouws. *Leksikografie*. Academica, Kaapstad, 1989.
6. J Markowitz, T Ahlswede and M Evens. 'Semantically Significant Patterns in Dictionary Definitions'. In *Proceedings of the Association of Computational Linguistics*, 1986.

7. R Morris. *LEXINET en die rekenarisering van taal (Hoofverslag van die LEXINET-program).* Raad vir Geesteswetenskaplike Navorsing, Pretoria, 1988.
8. J T Nutter and M W Evens. 'Building a Lexicon from Machine-Readable Dictionaries for Improved Information Retrieval'. *Literary and Linguistic Computing*, 5(2):129–138, (1990).

Acknowledgements: We thank the publishers of [3] for making the dictionary available for research. Without their kind permission this work would not have been possible.

Table 6. Afrikaans/English glossary of words appearing in examples.

Afrikaans	English
bakker	baker
bakkerswinkel	baker's shop
bakkery	bakery
bestuurder(s)	manager(s)
dier	animal
dignitaris	dignitary
diktator	dictator
diktatuur	dictatorship
diplomaat	diplomat
draak	dragon
draketand	dragon's tooth
eienaar(s)	owner(s)
evangelie	gospel
evangelis	evangelist
ewenaar	equator
flank	flank
gedeelte	part
gedeelte van 'n	part of a
iem.	somebody
middellyn	axial line
of	or
persoon	person
regering	government
skrywer	author
soort	sort / kind of
tand	tooth
toestel	apparatus / instrument
van	of
van 'n	of a
van die	of the
voet	foot
vrou	woman
wat	which / who / that / what

Notes for Contributors

The prime purpose of the journal is to publish original research papers in the fields of Computer Science and Information Systems, as well as shorter technical research papers. However, non-refereed review and exploratory articles of interest to the journal's readers will be considered for publication under sections marked as Communications or Viewpoints. While English is the preferred language of the journal, papers in Afrikaans will also be accepted. Typed manuscripts for review should be submitted in triplicate to the editor.

Form of Manuscript

Manuscripts for *review* should be prepared according to the following guidelines.

- Use wide margins and $1\frac{1}{2}$ or double spacing.
- The first page should include:
  - title (as brief as possible);
  - author's initials and surname;
  - author's affiliation and address;
  - an abstract of less than 200 words;
  - an appropriate keyword list;
  - a list of relevant Computing Review Categories.
- Tables and figures should be numbered and titled. Figures should be submitted as original line drawings/printouts, and not photocopies.
- References should be listed at the end of the text in alphabetic order of the (first) author's surname, and should be cited in the text in square brackets [1, 2, 3]. References should take the form shown at the end of these notes.

Manuscripts accepted for publication should comply with the above guidelines (except for the spacing requirements), and may be provided in one of the following formats (listed in order of preference):

1. As (a) LaTeX file(s), either on a diskette, or via e-mail/ftp – a LaTeX style file is available from the production editor;
2. In camera-ready format – a detailed page specification is available from the production editor;
3. As an ASCII file accompanied by a hard-copy showing formatting intentions:
   - Tables and figures should be on separate sheets of paper, clearly numbered on the back and ready for cutting and pasting. Figure titles should appear in the text where the figures are to be placed.
   - Mathematical and other symbols may be either handwritten or typed. Greek letters and unusual symbols should be identified in the margin, if they are not clear in the text.

   Further instructions on how to reduce page charges can be obtained from the production editor.
4. In a typed form, suitable for scanning.

Charges

Charges per final page will be levied on papers accepted for publication. They will be scaled to reflect scanning, typesetting, reproduction and other costs. Currently, the minimum rate is R20-00 per final page for LaTeX or camera-ready contributions and the maximum is R100-00 per page for contributions in typed format.

These charges may be waived upon request of the author and at the discretion of the editor.

Proofs

Proofs of accepted papers in categories 3 and 4 above will be sent to the author to ensure that typesetting is correct, and not for addition of new material or major amendments to the text. Corrected proofs should be returned to the production editor within three days.

Note that, in the case of camera-ready submissions, it is the author's responsibility to ensure that such submissions are error-free. However, the editor may recommend minor typesetting changes to be made before publication.

Letters and Communications

Letters to the editor are welcomed. They should be signed, and should be limited to less than about 500 words.

Announcements and communications of interest to the readership will be considered for publication in a separate section of the journal. Communications may also reflect minor research contributions. However, such communications will not be refereed and will not be deemed as fully-fledged publications for state subsidy purposes.

Book reviews

Contributions in this regard will be welcomed. Views and opinions expressed in such reviews should, however, be regarded as those of the reviewer alone.

Advertisement

Placement of advertisements at R1000-00 per full page per issue and R500-00 per half page per issue will be considered. These charges exclude specialized production costs which will be borne by the advertiser. Enquiries should be directed to the editor.

References

1. E Ashcroft and Z Manna. 'The translation of 'goto' programs to 'while' programs'. In *Proceedings of IFIP Congress 71*, pp. 250–255, Amsterdam, (1972). North-Holland.
2. C Bohm and G Jacopini. 'Flow diagrams, turing machines and languages with only two formation rules'. *Communications of the ACM*, 9:366–371, (1966).
3. S Ginsburg. *Mathematical theory of context free languages*. McGraw Hill, New York, 1966.

South African Computer Journal

Number 7, July 1992
ISSN 1015-7999

Suid-Afrikaanse Rekenaar-tydskrif

Nommer 7, Julie 1992
ISSN 1015-7999

Contents