Looking back: the first three years of the Lexicology Research Group*

Free University of Amsterdam, (September 1989–September 1992)

Willy MARTIN

1. Introduction

The title above is not meant as an announcement of a new theatrical play, nor as an adaptation of an existing one, rather, it is purely prosaic and down-to-earth. When being asked to write about our "travaux récents dans le domaine de l'informatique et de la statistique (dans les sciences humaines)", it seemed appropriate to me to take the opportunity to present what we have done up till now as a whole, thus providing a frame-work for the particular projects having been or being carried out. As a corollary then I will restrict myself to factual data of the projects in question, after having situated them in the general frame-work. What follows therefore will be a brief description of some computational projects situated within a particular research context with the intention to make a first exchange of ideas possible.

[*] The title of this article was inspired by a brochure distributed by the Knowledge-based Systems Group, University of Twente, announcing their first lustrum.

✉ Lexicology Subdepartment; Free University of Amsterdam (The Netherlands); 1105 De Boelelaan; 1081 HV Amsterdam.
Fax: + 31 20 661 30 54 E-mail: lexico@let.vu.nl

KEYWORDS: Lexicon, NLP-lexicon, lexicology, standardization of terminological description, machine-aided translation, indexing, information retrieval, re-usability of (bilingual) dictionaries, lexical knowledge base, parsing of definitions.

2. Research Approach

The approach to research taken by our group has been influenced by several factors. The major one being the fact that the Research Group developed from the teaching staff. In 1986 a Chair in Lexicology was founded at the Free University of Amsterdam, followed by the appointment of the undersigned. At the time only one collaborator was added, mainly for teaching purposes (terminology). The curriculum itself (a four-year-study) was called 'Lexicology' and had to integrate both theoretical and practical aspects of a study leading to such professions as lexicographer, lexical knowledge engineer and terminographer. The stance taken with regard to the lexicon in the curriculum development has had a direct bearing on the ensuing research. Summarizing and somewhat simplifying one could state that 'lexicon' as a research object, such as in teaching, has been interpreted:

not as "a set of existing words such as in a dictionary",

nor as "a set of existing idiosyncratic lexical items (morphemes and morpheme combinations)",

neither as "a set of existing idiosyncratic lexical items (morphemes and morpheme combinations) together with a set of morphological rules (to form possible non-idiosyncratic morpheme combinations)",

but as "an organized lexical knowledge bank needed by users so to be able to understand and produce language".

Next to this "internal" factor, a quite different, "external" one, has contributed to our group's research agenda. The fact viz. that the demand for *computerlexicons* became more and more stringent in the 80's, stressing such aspects as *extension* (the fact that computerlexicons should no longer be "toy" lexicons but able to cope with real life situations and so be large), *intelligence* (the fact that computerlexicons, next to knowledge w.r.t. "concrete" phenomena such as orthography and pronunciation, should also contain more "abstract" phenomena such as morphology, syntax and semantics/pragmatics), and *re-usability* (the fact that computerlexicons should be constructed in such a way that they can serve various users and lead to various applications).

From the above it can be inferred that our research is to be situated within the field of *computational lexicology / lexicography*, implying that

- *the lexical component of a natural language processing system* has played/plays a *central* role in it;
- such a component is seen as a *lexical knowledge base* entailing such aspects as *lexical knowledge acquisition, lexical knowledge representation* and *lexical knowledge application*;

- all these aspects are taken to contribute to the study of *language* in general, that of the (mental) *lexicon* in particular.

In what follows then the major research projects of our group will be presented, distributed over the three areas mentioned above, viz.: lexical knowledge acquisition/representation/application.

3. Research Projects

3.1. SNIV

The SNIV-project (Subcategorisatie-Nederlands-IBM-VU), sponsored by *IBM* The Netherlands, was directed at the construction of a basic NLP-lexicon for Dutch, containing some 25,000 lexical items and featuring the main characteristics of these items.

A lexical item (LI) was defined as a lexical meaning unit, its description being organized as a frame containing nine subframes. The subframes, as a rule, are sets of attribute-value pairs (in which values can be either atomic or complex). The "main" frame of an LI contains the following slots:

- part-of-speech,
- graphemics,
- morphology,
- syntax,
- syntactico-semantics,
- pragmatics,
- conceptual meaning,
- synonyms,
- usage.

The subframes may differ according to word-categories. The syntax-slot for nouns e.g. contains such features as:

- gender,
- number,
- number concordance,
- article,
- complementation.

The feature "article" in Dutch can then have the values: "de, het, de/het, noart". Although in this project the main stress obviously was put on the *lexical representation* facet, both *acquisition* and *application* aspects too were taken into account. So e.g. the constructed lexicon should function within a grammar and style checker for Dutch (under development with *IBM*), whereas w.r.t. acquisition a system called LEXCES (Lexical Constructing and Editing System) was built. This system was meant to restrict manual input to a minimum and expand and check it on the basis of rules (based on lexical relations). Starting from an input-line such as the one below [for the LI "kat" (= "cat")]:

```
SPE=kat GEN=M/F SEL=NONHUM SJF=GENDIERK SYN= 'poes'A1
  SUB=SUB huisdier SIZ=SIZ klein FU1=FU1 jagen
  A21=A21 (op) muis [*en] (en) vogel[*s]
```

LEXCES produced the following output

```
POS          CAT           NOUN
DESCR        DESCR         A1
GRAPH        SPELLING      kat
GRAPH        SPELVAR       SPELVARNIL
GRAPH        FORMVAR       FORMVARNIL
GRAPH        STATVAR       STATVARNIL
GRAPH        LENGTH        FULL
MORPH        MORPHSTR      SIMPMORPH
SYNT         GENDER        M/F
SYNT         NUMBER        LEXSING
SYNT         ARTICLE       DE
SYNT         NUMBERCO      NORMCONC
SYNT         COMPLEM       NCOMPLEM
SYNTSEM      REFER         COMMON
SYNTSEM      COUNTAB       COUNT
SYNTSEM      SELRESTR      NONHUMAN
PRAGM        ORIGIN        NORIG
PRAGM        STYLE         NSTYLE
PRAGM        CONNOT        NCONNOT
PRAGM        SOCIAL        GENSOC
PRAGM        SUBJF         GENDIERK
PRAGM        CHRONOL       SYNCHR
PRAGM        GEOGR         GENGEO
PRAGM        FREQ          BASIC
SYNON        SYNON         'poes' A1
USAGE        USAGE         (leeg)
CONCMEAN     SUPERTYPE     dier
CONCMEAN     SUBTYPE       huisdier
CONCMEAN     FUNCTION1     jagen
CONCMEAN     ARG21         (op)muis[*en] (en) vogel[*s]
CONCMEAN     SIZE          klein
ADM          SOURCE        BVD
ADM          COMMENT       (leeg)
```

The actual production of the lexicon was finished by February 1992. At the moment the data are loaded into the INGRES relational database system.

SNIV has been carried out by Hanny Demeersseman, Onno Huber, Willy Martin, Maurice Vliegen and Liesbet Willekens.

3.2. Eurotra-7

The Eurotra-7-project has been carried out under the auspices of the European Community by a consortium consisting of both academic and industrial partners during the period July 1990–June 1991. The task was to provide for a *feasibility* and *definition study* on the *re-usability* of lexical and terminological resources in computerized applications. The project gave rise to some 11 internal documents such as: Possible Applications of Re-usable Lexical Resources (Doc. 1); An Architecture for Re-usable Lexical Resources in a Multi-Theoretical Environment (Doc. 8); Feasibility of Standards for Collocational Description of Lexical Items (as a sub-document of Doc. 9, the syntax document); and Feasibility of Standards for Terminological Description of Lexical Items (Doc. 11).

Our group, represented by Willy Martin, Elsemiek ten Pas, Ilse Posch and Anke Schwarz, has mainly been involved in the work on (the definition and re-usability of) the representation of collocations and terms in collaboration with the University of Stuttgart (Institut für maschinelle Sprachverarbeitung) and Umist (University of Manchester, Institute of Science and Technology, Dept. of Computational Linguistics).

The ET-7 documents have set the pace for the EC's LRE (Linguistic Research and Engineering)-programme 1992–1994 and some other important lexical NLP-projects. Some of the reports will be published in the series "Studies in Machine Translation and Natural Language Processing" (to appear in 1993).

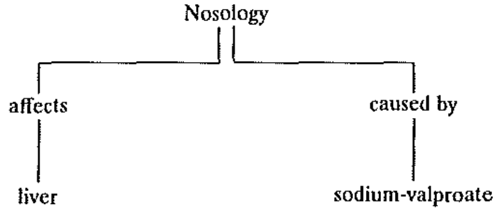
3.3. Sapiens

Sapiens is a one-year pilot project which has been carried out in collaboration with the University of Twente (Knowledge-based Systems Group) and sponsored by Elsevier Science Publishers at Amsterdam. Sapiens was directed at the *automatic generation of index terms for document abstracts*. As is known Elsevier publishes yearly the *Excerpta Medica* database, i.e. some 250,000 descriptions of primary medical literature (articles, books and the like) and adds to the document descriptions a number of index terms taken from a fixed list. The set of index terms is added manually and should give an indication of the contents of the original document. The user of the database thus can be helped when looking for relevant documents. In Sapiens an attempt has been made to design a system which automatically proposes index terms on the basis of an analysis of the title and the abstract of the document. Thereafter the human indexer could manually modify, reject or add index terms to the proposal as a last step in the process.

The tasks to be carried out by our group mainly involved the construction of a *lexicon* and a *conceptor*. At first sight the lexical data may seem much simpler than the ones to be found in SNIV. Yet for the first time a full semantico-conceptual description of lexemes (i.c. terms) was aimed at. This implied first the construction of a conceptual system to map the terms into. As a rule then a lexical entry is a three-place predicate in which the first argument is the lexeme to be described, the second argument is the concept-type it belongs to (plus, in the case of an "implicit" concept, the explicitation of the slots and, possibly, fillers of the concept in question), the third argument being the word category of the lexeme in question. Although I cannot elaborate here on the actual entries, an example may give one an idea of what is meant.

```
atlex (epilepsy, nosology-concept (
        [macro-concept (985321), funct-concept (231908)],
        [organism-concept (284623), unspecified, unspecified, unspecified]
        etc.                ),
        noun).
```

As a matter of fact the most important module we had to construct was the conceptor, a system which calculated, on the basis of wordcategorial and conceptual information, which strings in the text could be regarded as possible medical (index) terms. So e.g. from the title "A case of late onset sodium-valproate induced hepatitic failure" a structure will be calculated that contains a nosology concept that affects the "liver" and is caused by the drug "sodium-valproate". Or:



From this structure and from the information contained in the lexicon the following index terms were proposed (by means of selection and inference rules, devised by Twente): hepatitis, liver, sodium-valproate, sodium-valproate hepatitis, valproic-acid.

Although the project has ended in 1991 and has since then not been continued by Elsevier, it can safely be concluded from the results obtained, that this semi-automatic way shows considerable advantages over the current, fully manual one. Moreover our group has developed, in the wake of this project, *a concept-oriented parser of definitions* (of medical terms) which fully automatically parses definitions such as:

> asthma: "a respiratory disorder, often of allergic origin, characterized by difficulty in breathing, wheezing, and a sense of constriction in the chest"

into frames such as:

```
asthma:
   [ disease     caused-by          allergy]
   [ disease     g-affects          respiration]
   [ disease     has-symptom        constriction]
   [ disease     has-symptom        wheezing]
   [ disease     has-symptom        difficulty]
                 [constriction  g-affects   chest]
                 [difficulty    g-affects   breathing]
```

As one will have observed, starting from a NLP-project with a strong applicational bias (indexing, information retrieval) has again brought us back to the representation side in which structure and organization of lexical semantic data plays a predominant role. A substantial role has been played in this project by the following team members: Annelies Bon, Crit Cremers, Willy Martin, Mathilde Reedijk, Hennie van der Vliet and Liesbet Willekens.

3.4. Multilex

Multilex is an ESPRIT-project which has been started in 1991 and should end in 1993. As most of the projects sponsored by the EC several European institutes both from industry and academia take part in it. Our concern in this project has mainly been in the *actual construction of standards for the terminological description of lexical items, particularly in the domains of semantics, pragmatics and transfer.*

This nicely fits in both with teaching tasks (viz. terminology/terminography) and previous research undertaken (Eurotra-7, Sapiens). To give an idea of what is meant two examples are given: the first is a semantic description of the medical term "lung", the second is a BNF-notation of the transfer specifications needed for terms.

```
semantics: [
   conceptual_definition:
         def_text:       "One of a pair of highly elastic cone-shaped organs of
                         respiration occupying the thoracic cavity, where each is
                         surrounded by a pleural sac and separated from the other
                         by the heart and other contents of the mediasticum."
         def_source:     source type: DIC
                         : title: Churchill's illustrated medical
                         dictionary
                         : edition: 1
                         : place_of_publ.: New York, Edinburgh, London, Melbourne
                         : date_of_publ.: 1989

   conceptual_relations:
         isa             : organ_lu_x
         typ_number      : 2
         typ_attr        : elastic_lu_x, cone_shaped_lu_x
         funct           : respiration_lu_x
         loc_in          : thoracic_cavity_lu_x
         loc_ad          : pleural_sac_lu_x, heart_lu_x, mediasticum_lu_x

   conceptual_collocations:
         typ_attr_spec   : honeycomb_lung, hyperlucent_lung

   ostensive_definition  :        <pointer to picture>

   examples: "....","....","...."
                                  ]

TRANSFER     ::=     TARGET_LANGUAGE        TARGET_LU         LINK_STATUS
            VAR_STATUS      CONC_EQUITYPE      PRAGM_CONTR      {EXAMPLES}

TARGET_LANGUAGE        ::= "string"
LINK_STATUS            ::= lexicalized | semi-lexicalized | non-lexicalized
VAR_STATUS             ::= "<Boolean>"
CONC_EQUITYPE          ::= complete | PARTIAL | NON
PARTIAL                ::= interl_hypero [SEM_SPEC<interl_hypero>] |
                       interl_hypo CONSTR<interl_hypo> | related
NON                    ::= subst_by_explanation | subst_by_borrowing |
                       subst_by_near_equi
SEM_SPEC<interl_hypero>          ::="string"
CONSTR (interl_hypo)   ::= SEM_CONSTR | CONTEXT_CONSTR
SEM_CONSTR             ::= "string"
CONTEXT_CONSTR         ::= "string"
PRAGM_CONTR            ::= nil | [diff_origin] [diff_comm_sit]...

EXAMPLES               ::= SOURCE_EX (TARGET_TRANS)
SOURCE_EX              ::= "string"
TARGET_TRANS           ::= "string"
```

Although both examples once more stress *representation aspects* and so e.g. should be taken into account when constructing *termbanks*, Multilex should ultimately lead to a *multilingual, multipurpose lexical database* and so facilitate different types of NLP-processing. Researchers active in this project have been/are: Hanny Demeersseman, Willy Martin, Elsemiek ten Pas, Jérôme Paul and Maurice Vliegen.

3.5. Wordtranslator

Whereas most projects mentioned up till now focussed on representational aspects of NLP-lexicons, and, as such, aimed to be multifunctional, this project has but one clearly delineated objective: *to disambiguate English texts at wordlevel and translate the thus disambiguated items into Dutch.*

In this respect *Wordtranslator* takes for granted the knowledge to be found in a bilingual dictionary (Van Dale's Unabridged Dictionary English–Dutch) and adds to the *static knowledge* found there (declarative knowledge, knowledge of the "what"-type), *dynamic knowledge* (procedural knowledge, knowledge of the "how"-type). On the basis of the two then *Wordtranslator* will (try to) select the translation equivalent which is most adequate in the context at stake.

Suppose that this context reads: "He was not motivated enough to make a pile" and that one does not know how to translate "pile" into Dutch, then *Wordtranslator* will first lemmatize and tag the text categorically (and so it will observe that "pile" e.g. is a noun, not a verb). Then it will consult, based on that knowledge, the "nominal" readings and translations of "pile" the Van Dale English–Dutch databank contains, from which it will learn that "pile" has both countable and uncountable readings. Therefore it will try to prove that "pile" is either countable or uncountable (it will succeed in proving the countability of "pile" in the case at stake), etc. All this happens in a real-time mode, ending up with the following output:[1]

```
... he was not sufficiently motivated to make a PILE.

Subcategories before det-subcat: count; uncount
Subcategories after det-subcat: count

Pile had 14 main translations and six examples.
There is a contextmatch. In this context PILE can be translated as
hoop/berg geld VAR fortuin
make a/one's pile: fortuin maken
```

[1] For the sake of clarity the Dutch commentary has been translated into English here.

As *Wordtranslator* is meant to be an intelligent electronic lexical translation aid it tackles the lexicon now from the application-side and forces it, so to speak, to contain information which can be used to disambiguate. In this respect it is particularly interesting to see how "contextual" information is to be defined and organized.

Wordtranslator has been developed by Annelies Bon under the supervision of Willy Martin during the academic year 1991–1992 for Van Dale Publishers, Utrecht, The Netherlands. This project will be continued.

3.6. Feasibility Study on the Re-usability of Bilingual Dictionaries

Actually the project under review is somewhat more restricted as the title above might suggest. The project, sponsored by Van Dale Publishers, Utrecht, had to investigate the feasibility of deriving from a series of bilingual dictionaries in which Dutch figured as either a source or as a target language, a new series of bilingual dictionaries which linked foreign languages (from the Dutch speaker's point-of-view) to each other. This small scale project (carried out by Hennie van der Vliet under the supervision of Willy Martin during the first half of the academic year 1991–1992) has led to a programme which takes French–Dutch dictionary text as an input and, via Dutch–French and Dutch–German, fully automatically delivers output such as the following:

Fabrikat
0.1 produit ⟨m⟩ **manufactur/-e** perfect match
0.2 fabrication ⟨v⟩ ⇒ marque, [mbt. wijn] cru ⟨m⟩ perfect match
♦ 2.1 ⟨inf.⟩[scherts] dieser Wein ist eigenes -
(2).2 Schuhe von slechtem Fabrikat, von schlechter Qualität
(2).2 niederländisches Erzeugnis, Fabrikat *(produit de) fabrication néerlandaise*
(2).2 dieser Wein ist eigenes Fabrikat?? *produit de la maison*??
(2).1 ein Fertigerzeugnis, Ganzfabrikat

the remainder:
[constructie, vorm] **fabrication** ⟨v.⟩ ⇒ façon ⟨v.⟩
[produkt] **produit** ⟨m.⟩ ⇒ production ⟨v.⟩
♦ (2).? dieser Pudding ist eigenes Fabrikat *c'est une crème de fabrication maison*

From a practical point-of-view the results prove to be both promising and tractable. From a lexical-theoretical point-of-view, the project has been interesting in that it could shed some new light on the notion of an "interlingua"-lexicon (in our case Dutch) and on the operational definition of the conditions that the latter should obey to so for it to function properly (a.o. involving such aspects as synonymy, symmetry, transitivity and equivalence). Most probably this project will lead to a full scale exploitation by Van Dale Publishers.

3.7. Current projects

At the moment of writing this text some projects are being carried out for which it is still too premature to give any detailed results. For the sake of completeness I will restrict myself to summing the latter up:

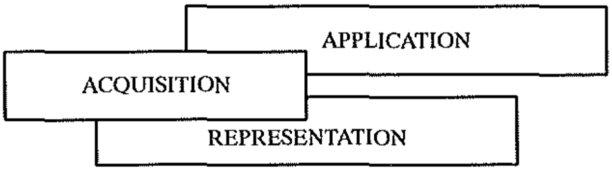
- D-TALE (Dutch Tagger-Lemmatizer)
- Semantic Tagger (Tagger meant to enrich the lexicon first – using a mixture of both a rule-based and a data-oriented approach – and thereafter texts)
- Delis (LRE-project: Descriptive lexical specifications and tools for corpus-based lexicon building)
- Automatic Corpus-exploration (within Acquilex-II, a BR-ESPRIT project).

4. Conclusion

It should be clear form the above that what has been presented here is *team work*: it is, of course, not possible for an individual to carry out so many projects in so short a lapse of time. At the moment our group consists of about eight researchers (not all of them working fulltime though). What brings projects and researchers together is the interest taken in the NLP-lexicon: the lexicon as a central component of an NLP-system. As stated already this entails with it both theoretical and practical problem solving distributed over the areas of

- acquisition,
- representation,
- and application.

As we see it there is not always a clearcut division between these three areas, instead they show some overlap as the figure underneath may illustrate:



As will be noticed, each of the three aspects is linked with the other two, though it remains true that representation is central to the whole of it and a.o. therefore shows the greatest amount of overlap. It is to be hoped that in the years to come we will be able to keep the NLP-lexicon in the focus of our attention and explore further its manifold aspects and interrelationships.

5. References

All the projects are documented by internal reports. I will not mention them here, those interested can contact the undersigned, some of them are public domain, for some others some (small) charge has to be paid. Instead of mentioning the reports for each project one relevant publication (mostly article) will be mentioned as an addition to the information given in section 3 of this article:

- ad SNIV: H. DEMEERSSEMAN, "The construction of a kernel lexicon of Dutch" [in Dutch], in *Jaarboek Corpusgebaseerde Woordanalyse*, VU Amsterdam, 1992, 25–42.
- ad Eurotra-7: U. HEID and J. MCNAUGHT, *Eurotra-7 Study, Final Report*, Stuttgart, 1991.
- ad Sapiens: W. MARTIN, "(More) Intelligent Information Retrieval: the Parsing of Definitions" [in Dutch], in *Jaarboek Corpusgebaseerde Woordanalyse*, VU Amsterdam, 1992, 99–111.
- ad Multilex: only internal reports published.
- ad Wordtranslator: W. MARTIN, "Wordtranslator: a semi-automatic lexical system for the translation of texts", to appear in *Toegepaste Taalwetenschap in Artikelen*.
- ad Feasibility Study on Re-usability: W. MARTIN and H. VAN DER VLIET, "Semi-Automatic Generation of a Bilingual Dictionary" [in Dutch], to appear in: *Proceedings of STINFON92 (Stichting Informatiewetenschappen), 18 December 1992.*