Lexical Database Design: The *Shakespeare Dictionary* Model

H. Joachim Neuhaus

Westfälische Wilhelms–Universität, FB 12
D–4400 Münster, West Germany

1. The Data

The *Shakespeare Dictionary* (SHAD) project has been using structured databases since 1983. The system is implemented on a PRIME 250–II computer using standard CODASYL–DBMS software and related tools. The project has been able to draw on a vast repository of computerized material dealing with Shakespeare and the English lexicon. Initially, it was part of the "Sonderforschungsbereich 100 Elektronische Sprachforschung" sponsored on the national level by the Deutsche Forschungsgemeinschaft. The research team has been directed by Marvin Spevack and H. Joachim Neuhaus, now both at Münster, and Thomas Finkenstaedt, now at Augsburg. Spevack's *Complete and Systematic Concordance to the Works of Shakespeare* (Hildesheim and New York, 1968–1978) and Finkenstaedt's *Chronological English Dictionary* (Heidelberg, 1970), both in machine readable form, were used in a computer–assisted lemmatization procedure (Spevack, Neuhaus, and Finken–staedt 1974).

A chronologically arranged dictionary, where entries are sorted according to the year of first occurrence, makes it possible to "stop" the development of the recorded English vocabulary at any desired moment and to compare, for instance, Shakespeare's vocabulary with the corpus of English words recorded up to 1623, when the *First Folio* appeared (Neuhaus 1978). The set of words in Shakespeare can be compared with the complement set of words available in Elizabethan English, but not attested in Shakespeare's works. In this way there is a systematic integration into the total vocabulary. As a result, our database model can easily be expanded or transfered to cover larger or different vocabularies.

In order to present the complete Shakespearean vocabulary and to disengage SHAD from dependence on a single edition of Shakespeare, the data were expanded to include all stage directions and speech–prefixes in all quartos up to and including the *First Folio* (Volume VII of the *Complete and Systematic Concordance to the Works of Shakespeare*), and the "bad" quartos (Volume VIII). Volume IX presents all substantive variants, producing a composite Shakespearean vocabulary in modern and eventually old spelling.

In analysing this material a strict differentiation between vocabulary level and text level has been observed. Further data–preparation on the vocabulary level concentrated on formal properties of Shakespearean lemmata, such as morphological structure, or etymological background. There is a complete morphology for all lemmata (ca. 20,000 records), which gives detailed structural descriptions of derivations, compounds, and other combinations, as well as all inflected word–forms, as they occur in the text. The etymological data include word histories and loan relations, again supplemented by chronological data. Content–oriented criteria were used in a taxonomic classification of all lemmata (Spevack 1977). On the whole, there are more than thirty fields of information in the original lemma–record file. For the multidimensional analysis and presentation of these resources it seemed natural to use database concepts.

Due to a special intervention of the Deutsche Forschungsgemeinschaft and the support of the Ministry, which we both gratefully acknowledge, we could implement our first database in 1983 on a newly installed PRIME 250–II computer. The PRIME DBMS software, which we use, is actually one of the first commercial products which closely adhered to the CODASYL network data model. The design started with a database schema for Shakespearean word–formation and etymology. Since then the system has grown steadily including now a thesaurus structure and a link to the text itself. The database is accessed in batch mode using the FORTRAN and COBOL interfaces, and interactively with the VISTA query language and report generator. Of course, in a first implementation not only the database schema itself, but the preparation of files, and the programming of the database creation job have to be carried out. The first word–formation database was established in three separate steps. The total time needed to complete the job was about 17 hours. Physical design is especially important in large databases. Our Münster team was interested in that aspect from the very beginning (Döge 1984).

2. Preliminary Design Considerations

Linguists and lexicographers are latecomers to the field of database applications. Database software has been available since the early 1960's. The early 1970's brought a wide variety of commercial products and a consolidation on the conceptual side, which ultimately led to standardization, design philosophies, and specifications of "normal forms". At that time lexicographers still used the concept of an archive when talking about new technologies, such as Barnart (1973), Chapman (1973), and Lehmann (1973) at the 1972 *International Conference on Lexicography in English.* Similarly, in the late 1970's, we witnessed preparations for a Stanford Computer Archive of Language Materials. There is nothing wrong with the idea of an archive. But a database is

something different. By now, the expression "database" should only be used as a technical term. Perhaps "data bank" may be used instead of "database" when talking about files of data, or archives in a conventional sense. The *Association for Literary and Linguistic Computing* may have had this clarification in mind when naming its specialist group "Structured Data Bases".

Although hierarchical data models and network models had been available since the early 1960s, and relational architectures since the early 1970s (Codd 1970), software implementations were not generally accessible in university computing centres due to high cost, and lack of special support. Although the Münster computing centre had the hierarchical IMS software, a product of IBM, it was not made available for our project. Looking back from today, that may not have been a handicap for at least two reasons: lexical relationships are only rarely hierarchical in a natural sense, and, more importantly, hierarchical systems do not have a common standard. There is no migration path from one software product to another. Since a Shakespeare database will have a rather long life cycle, and was meant to be a model for similar projects, the requirement of a standard model seemed to be imperative. The process of standardization has been proceeding more rapidly for the CODASYL network model than for any other architecture. In the early 1980s there was just this model that fulfilled our requirements, and this is basically true even today.

Beginning with the early 1980's lexical symposia and conferences had an ample share of papers reporting on ongoing research which used the database concept in a variety of ways. In 1981 Nagao et al. reported on "An Attempt to Computerize Dictionary Data Bases" (1982). At the same conference a University of Bonn group (Brustkern and Hess 1982) presented "The Bonnlex Lexicon System", which two years later evolved into a "Cumulated Word Data Base for the German Language" (Brustkern and Schulze 1983). A list of similar projects could easily be extended. One might have expected that the logical design of lexical databases would have built on structural linguistics, where we typically find entities and relationships, and in general, set theoretic notions, which can directly be translated into conceptual data-structures.

Surprisingly, in many designs, linguistic considerations did not seem to have played a major role. Instead, the authors simulate conventional lay-out and typesetting arrangements of printed dictionaries. An example is the widespread dictionary usage to print one "Headword" in bold type and then use special symbols, such as the tilde, to refer to the headword, or parts of it, thus saving space for the treatment of further lexical items with the same spelling. Nagao et al. (1982) very faithfully transfered this and other lay-out details into their design. But should a conventional "Headword" and its dependencies be a serious candidate for a database entity? Are the reasons that led dictionary publishers to accept certain lay-out techniques at all relevant for an electronic database? These questions seem not to have been raised. The design seems to have

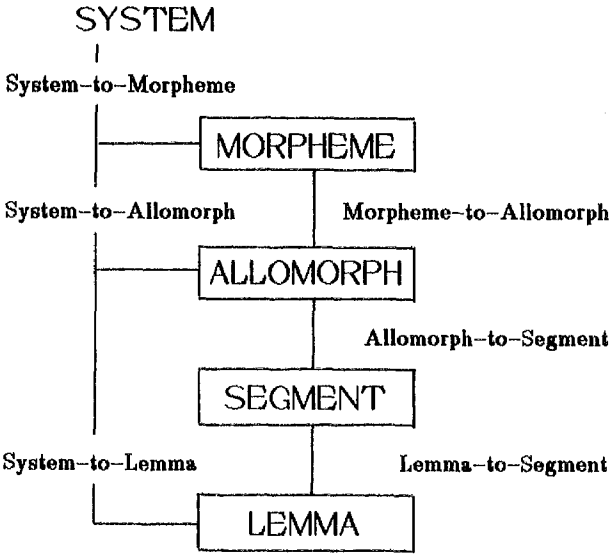

Figure 1. Data-Structure for Morphological Families
(SHAD, database fragment)

become a paradigm case of an imitation design, where a new technology replicates design features of an older technology. The basic misunderstanding is the false identification of a mere presentation in a printed dictionary with an underlying lexical information structure.

If the "Headword" is not a relevant database entity, which entity should be taken instead? There is only one serious candidate: the lemma. The lemma is a well defined linguistic notion. It is also well known in computational work due to various automatic or semi-automatic lemmatization algorithms. It is an abstract notion in the sense that printed dictionaries and database systems need a lemma-name to refer to it. Language specific conventions usually govern the choice of a lemma-name. Latin verbs, for example, are customarily lemmatized using the first person singular present form as lemma-name. A lemma is the set of all its inflected word-forms. It thus comprises a complete inflectional paradigm. Some lemmata have defective paradigms or suppletive paradigms. Conventional dictionaries quite often include paradigmatic information in their front matter. The user has to relate specific cases to these examples. A database can relate these explicitly. A natural way to do this is by a one-to-many relationship between lemma and word-form. In an author dictionary word-forms will be further related to the text, and its internal structure.

A machine-readable dictionary is just a starting point for a structured lexical database. In the Bonn "Word Data Base for the German Language" (Brustkern and Schulze 1983b) there is but one database entity, "Lexical Entry", which seems to correspond to the lemma rather than to a "Headword". The authors speak about the "microstructure" and the "macrostructure" in respect to

"Lexical Entries", but only the former is discussed in detail. The later is only mentioned once: "Special characteristics of the macrostructure (other than alphabetical order) are to be made explicit in the logical structure of the data base" (Brustkern and Schulze 1983b). "Macrostructure" is rarely visible in a conventional alphabetic dictionary, although we are used to "synonyms" and "antonyms", dictionary "senses", and labels that identify technical jargon, or special terminologies in individual dictionary entries. In the design of a lexical database it is useful to make these various relations between lemmata explicit. In this manner a user gets more information than by consulting a printed dictionary. The information he gets is related and structured in unexpected ways.

3. A Sample Schema

There are various ways to approach the problem of schema design. For the *Shakespeare Dictionary* Morphology Database, now an integrated part of the overall architecture, both object-class methods and query-assertion methods lead to the current schema (cf. Figure 1). There are four base object-classes (entities): *lemmata, segments, allomorphs,* and *morphemes* having cardinality values between 2,500 and 40,000 records. Queries were to allow for a direct retrieval on three levels: the conventional level of the lemma, the level of allomorphs, and the morphemic level. This is achieved by a virtual record, defined as a subschema (cf. Figure 2). In this way the database design mirrors a structural morphological analysis directly. The concept of a *morphological family* defined as a set of lemmata which has at least one morpheme in common is thus immediately accessible for database queries.

The ultimately Latin prefix { IN– } has, for example, database links to allomorphs such as { **im–** } in the lemma **impure**, { **il–** } in the lemma **illegitimate**, or { **ir–** } in the lemma **irregular**. In Shakespeare's vocabulary there are almost 200 lemmata which belong to this { IN– } family. A statistical survey of morphological families in Shakespeare, reveals characteristic "family types". Since morphological descriptions are directly accessible for a study of patterns such as nominal compounds, conversions, or derivations, listings of morphologically similar lemmata supplement family

```
/*-------------------------------------*/
VIRTUAL RECORD SECTION.
   VIRTUAL RECORD MORPHEME-TO-LEMMA;
      BASE RECORD IS SEGMENT;
         MORPHEME OWNS ALLOMORPH
           VIA MORPHEME-TO-ALLOMORPH;
         ALLOMORPH OWNS SEGMENT
           VIA ALLOMORPH-TO-SEGMENT;
         LEMMA OWNS SEGMENT
           VIA LEMMA-TO-SEGMENT.
/*-------------------------------------*/
```

Figure 2. Virtual Record for Morphological Families
SHAD database subschema

		Frequency	Dating	
Morpheme	{ SPEAK }			
Allomorph	{ **speak** }			
vb.	**speak**	111	Oldeng.	
	bespeak	13	Oldeng.	
	mis–speak	1	1200	
	forspeak	1	1300	
n.	**speaker**	11	1303	
vb.	**unspeak**	4	1340	
adj.	**unspeakable**	5	1400	
pp.	**false–speaking**	2	1593	SON
vb.	**respeak**	1	1600	HAM
	outspeak	1	1603	
Allomorph	{ **spok–** }			
pp.	**well–spoken**	3	1400	
	fair–spoken	1	1460	
n.	**spokesman**	1	1540	
pp.	**foul–spoken**	1	1593	TIT
Allomorph	{ **speech** }			
n.	**speech**	159	Oldeng.	
adj.	**speechless**	15	Oldeng.	

Shakespeare Datings:
HAM *Hamlet,* SON *Sonnets,* TIT *Titus Andronicus*

Figure 3. A Morphological Family in Shakespeare's Vocabulary

listings in a study of the morphological articulation of Shakespeare's vocabulary. The database has access to various additional and specialized kinds of morphological information such sound symbolism, popular etymology, or contamination. Furthermore, morphological information is by design linked with etymological information. Morphological families which are etymologically related can be grouped together under one etymon. One example for such an etymological grouping is given in Figure 4. The phenomenon of etymologically homogeneous or disparate word-formation, which has traditionally been of some interest in Shakespearean studies can be analysed directly. These materials are currently being prepared for the forthcoming first volume of SHAD.

Any lexical database design should account for external links with other lexical databases (Neuhaus 1985). Here again, a common standard is essential. The *lemma record* is a natural interface in these external relations. Standardization of the lemma concept may therefore be a first step for systematic database connections.

		Frequency	Dating
Family	**trow**		
vb.	**trow**	17	Oldeng.
n.	**troth**	111	1175
vb.	**betroth**	12	1303
adj.	**troth–plight**	2	1330
n.	**troth–plight**	1	1513
pp.	**new–trothed**	1	1598
pp.	**fair–betrothed**	1	1607
Family	**truce**		
n.	**truce**	15	1225
Family	**true**		
adj.	**true**	849	Oldeng.
adv.	**truly**	180	Oldeng.
n.	**truth**	361	Oldeng.
adj.	**untrue**	7	Oldeng.
n.	**untruth**	4	Oldeng.
n.	**true–love**	10	800
n.	**true**	36	1300
pp.	**true–hearted**	3	1471
pp.	**truer–hearted**	1	1471
n.	**truepenny**	1	1519
pp.	**true–born**	2	1589
pp.	**true–anointed**	1	1590
pp.	**true–derived**	1	1592
pp.	**true–disposing**	1	1592
pp.	**true–divining**	1	1593
pp.	**true–telling**	1	1593
pp.	**true–devoted**	1	1594
adj.	**honest–true**	1	1596
pp.	**true–begotten**	1	1596
pp.	**true–bred**	3	1596
pp.	**true–fixed**	1	1599
pp.	**true–meant**	1	1604
Family	**trust**		
n.	**trust**	1	1225
vb.	**trust**	22	1225
adj.	**trusty**	21	1225
n.	**mistrust**	9	1374
vb.	**mistrust**	14	1374
vb.	**distrust**	3	1430
n.	**distrust**	3	1513
adj.	**mistrustful**	2	1529
adj.	**trustless**	1	1530
n.	**truster**	2	1537
n.	**self–trust**	1	1583
adj.	**distrustful**	1	1589

Figure 4. Etymological Grouping of four Morphological Families

References

Barnhart, Clarence L. "Plan for a Central Archive for Lexicography in English." In *Annals of the New York Academy of Sciences,* No. 211 (1975), pp. 302–306.

Brustkern, J. and K. H. Hess. "The Bonnlex Lexicon System." In *Lexicography in the Electronic Age.* Ed. J Goetschalckx and L. Rolling. Amsterdam: North–Holland, 1982, pp. 33–40.

Brustkern, J. and W. Schulze. "Towards a Cumulated Word Data Base for the German Language." Proc. Sixth International Conference on Computers and the Humanities. 6–8 June 1983. Raleigh, North Carolina.

"The Structure of the Word Data Base for the German Language." Proc. International Conference on Data Bases in the Humanities and Social Sciences. 10–12 June 1983. New Brunswick, New Jersey.

Chapman, Robert L. "On Collecting for the Central Archive." In *Annals of the New York Academy of Sciences,* No. 211 (1975), pp. 307–311.

Codd, E. F. "A Relational Model of Data for Large Shared Data Banks," *Communications of the ACM,* 13.6 (1970), 377–387.

Döge, Michael. *Probleme eines CODASYL–Datenbank–systems, dargestellt am Beispiel des DBMS–Software–Paketes der Firma PRIME.* Münster, 1984.

Finkenstaedt, Thomas. *A Chronological English Dictionary. Listing 80,000 Words in Order of their Earliest Known Occurrence.* Heidelberg, 1970 (with Ernst Leisi, Dieter Wolff).

Lehmann W. P. "On the Design of a Central Archive for Lexicography in English." In *Annals of the New York Academy of Sciences,* No. 211 (1975), pp. 312–317.

Nagao, M. et al. "An Attempt to Computerize Dictionary Data Bases." In *Lexicography in the Electronic Age.* Ed. J Goetschalckx and L. Rolling. Amsterdam: North–Holland, 1982, pp. 51–73.

Neuhaus, H. Joachim. "Author Vocabularies compared with Chronological Dictionaries." *Bulletin of the Association for Literary and Linguistic Computing,* 6 (1978) 15–19.

"Design Options for a Lexical Database of Old English." *Problems of Old English Lexicography.* Ed. Alfred Bammesberger. Eichstätter Beiträge 115. Regensburg, 1985, 197–210.

Spevack, Marvin. *A Complete and Systematic Concordance to the Works of Shakespeare.* 9 volumes. Hildesheim, 1968–1980.

"SHAD: A Shakespeare Dictionary," *Computers in the Humanities.* Ed. J. L. Mitchell. Edinburgh, 1974, 111–123. (with Th. Finkenstaedt, H. J. Neuhaus)

"SHAD (A Shakespeare Dictionary). Toward a Taxonomic Classification of the Shakespeare Corpus." *Computing in the Humanities. Proceedings of the Third International Conference on Computing in the Humanities.* Ed. Serge Lusignan and John S. North. Waterloo, Ontario, 1977, 107–114.