A Corpus-based Electronic Dictionary for (Re)search

Hanne Ruus
University of Copenhagen, Institute of Nordic Philology
Njalsgade 80
DK-2300 Copenhagen S
Denmark
haru@hum.ku.dk

Abstract
The paper describes an electronic dictionary recently published on cd-rom accompanied by the text corpus on which it is based. The dictionary is a key to the corpus consisting of the oldest collections of Mediaeval Danish ballads, nine manuscripts from the latter half of the 16th century. The dictionary is constructed to function as a search tool for the interested layman, at the same time offering numerous possibilities for linguistic and literary studies. The computational philological methodology subdivides the task of constructing the dictionary into eight processes, five of which are fully automatic. The accumulating approach takes advantage of the information in all the material processed at any given time and makes the current dictionary and the processed texts available for scholarly investigations during the dictionary construction process. The structure of the dictionary reflects the multi level markup of the texts. The use of the descriptive levels in the dictionary is illustrated by examples of research from several disciplines: discourse analysis, semantics, lexicology, and studies of style and genre.

Introduction

With the steady growth of computer power and the general availability of PCs, the literary heritage can be made accessible to the interested general public with a corpus-based dictionary as a key to a digital version of the authentic texts, illustrated by pictures of their manuscripts. In the Scandinavian literary heritage, the Danish ballad manuscripts of the 16th century occupy a prominent place as the oldest written evidence of the oral European ballad tradition. In a five years' interdisciplinary research project, *Dansk Folkevisekultur 1550-1700* (*Danish Balladry 1550-1700)*, the oldest manuscripts, the reception of the ballads, and their cultural context were the main themes of study [Ruus 1996]. The results are being published in four volumes [Lundgreen-Nielsen & Ruus 1999, 2000, 2001, 2002], of which the third is accompanied by a cd-rom [Ruus 2001b] giving access to a dictionary based on 547 ballad and song texts, and to the complete textual tradition before 1591, the year of publication of the first printed ballad edition.

The Textual Challenge

Most of the ballad and song texts have been published in scholarly editions grouped by genre, not by manuscript, and the only dictionary covering the relevant period of Danish was published before the scholarly editions were finished; there is consequently a need for a lexical key to the texts. The manuscripts are written by numerous different persons at a time when instruction in writing was scarce, and regularity and orthographical order was considered less important than the imposing effect of double and triple letters. In the manuscripts, words like *heart* and *heaven*,

Modern Danish *hjerte* and *himmerig*, are spelled in 29 and 36 different ways respectively, Figure 1 and Figure 2 show extracts from the entries in the dictionary. In order to search for and find all instances of a given word or phrase in the text corpus, it is necessary to neutralize such orthographic variation; even the most learned philologist will be unable to foresee all the spellings of a given word or phrase in the manuscripts. The neutralization is effected by adding a neutral spelling as close to the spelling in Modern Danish as the metre permits, to each textual word. To facilitate searches for all forms of a given lemma, the neutral forms are tied to the relevant lemmas. In this way the texts are represented on three different levels: the first level being textual source word and abbreviated parts of the source word, the second an orthographically neutral spelling, the third the corresponding lemma with its part of speech. The three levels in the texts were marked up according to the SGML standard.

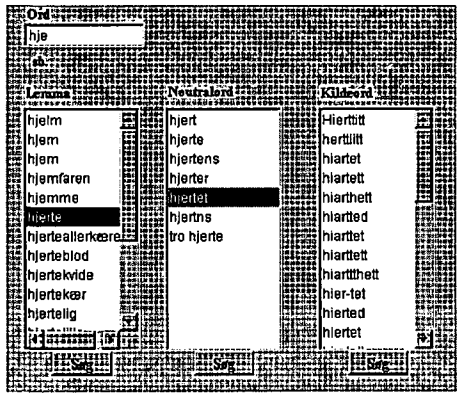


Figure 1: Some spellings of *hjertet.*

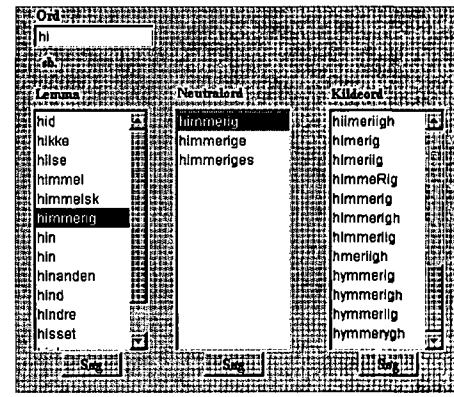
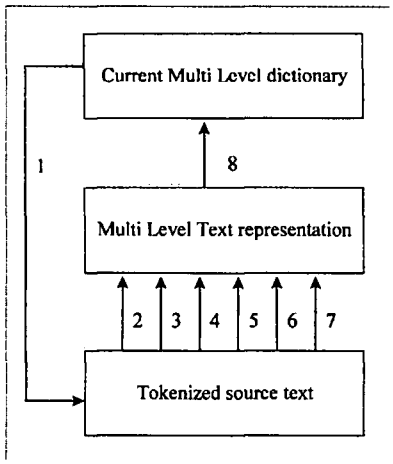


Figure 2: Some spellings of *himmerig.*

The Methodology for Constructing the Dictionary

The methodology for furnishing the texts with the markup, at the same time accumulatively building the dictionary, subdivides the task into 8 processes,[Duncker & Ruus 2000; Ruus 2001a], cf. Figure 3. The dictionary is built accumulatively by processing one text at a time. The first step in the processing of a text is a look-up of all the source word forms in the current dictionary, information about all the words already in the dictionary is added automatically to the source word forms in the text. The next step is a fully automatic process that adds proposals for neutral forms and lemma forms for the unknown word forms in the text that resemble forms of already known words. As the third step, the lexicographer fills in information about the remaining unknown word forms. At this stage, many words in the text have several proposals, these proposals are organized in wheels, and an automatic process based on the likelihood of the proposal in the current context, justifies the wheels so that the most likely proposal is presented first. In the next step, the lexicographer checks the multi level representation. As a further precautionary measure, a dictionary of new words in the text is made automatically and checked by the lexicographer. When a text has been processed the new source word forms, new orthographically neutral forms, and new lemmas from this text is added to the dictionary.



1. Add MLT information to known source word forms.
2. Forecast MLT information for unknown source word forms related to words in the MLT dictionary.
3. Fill in MLT information for remaining unknown source word forms.
4. Adjust homograph wheels.
5. Check Multi Level representation of the text.
6. Make dictionary of new words from the text.
7. Check dictionary of new words from the text.
8. Update MLT dictionary with the new words from the text.

Figure 3: Eight processes are combined to add multi level information to a text, extract its new words and add these to the current dictionary. Five of the processes are fully automatic exploiting the accuracy and speed of computers, three processes are performed under surveillance of the lexicographer because they require the unique hermeneutic powers of humans. The processes 1, 2, 4, 6, and 8 are fully automatic, the processes 3, 5, and 7 are supervised by the lexicographer.

The Dictionary Structure and Content

When the texts are being marked up, the unknown textual words and lemmas from each of the texts are added to the dictionary, all multi level information about the textual words is stored in the dictionary. The structure of the dictionary entries is as follows (the arrow means 'consists of', the question mark means 'one or none', the plus sign means 'one or more') :

dictionary entry ----> lemma, part of speech, discriminator?, gloss?,
(orthographically neutral spelling, (source word)+)+

The lemmas are, as far as possible, chosen in accordance with the lemmalist of the authorised Modern Danish dictionary of orthography [Dansk Sprognævn 1986], two or more homographic lemmas with the same part of speech are distinguished by discriminators. Lemmas completely foreign to Modern Danes in form or meaning receive a gloss. The dictionary extracts in Figure 4 and 5 are taken from four entries distinguished pairwise by discriminators: the verbs *bede vb; anmode* (En. *ask*) and *bede vb; jage* (En. *hunt*), and the nouns *ting sb; forsamling* (En. *court*) and *ting sb; genstand* (En. *thing*). In the extracts (cf. Figure 1, 2, 4, 5 and 9), the discriminator and part of speech of the lemma marked in the list to the left, appear above, the neutral forms of the marked lemma in the middle, and source forms of the marked neutral form to the right.

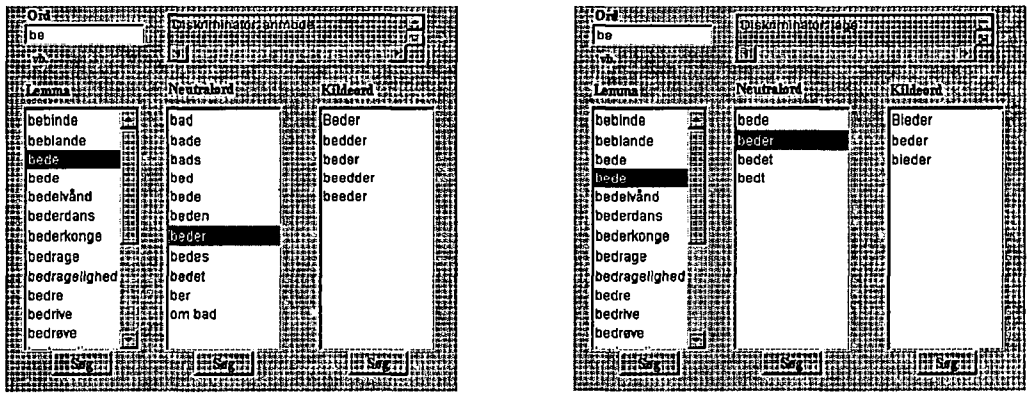


Figure 4: Look-ups of the two verbs *bede*; the lemmalist is in the column to the left, the orthographically neutral forms of the selected lemma appear in the column in the middle, and the source forms of the selected neutral form in the column to the right.

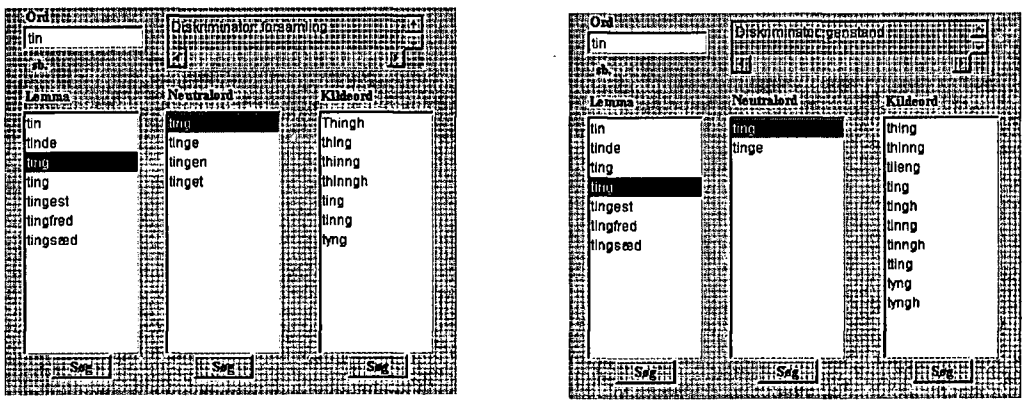


Figure 5: Look-ups of the two nouns *ting*. Note the similarities in the two lists of source forms in the columns to the right. When the button marked **Søg** is clicked, the relevant information is entered automatically into a search window.

The Pivotal Role of the Dictionary

The cd-rom has a user interface with buttons, windows, and pull-down menus. The user may search in the whole corpus or restrict the search basis according to his interests, e.g. to a single manuscript or to all the versions of one single ballad. When the dictionary button is clicked, a window with a box for indicating the relevant part of the dictionary appears, when a few letters are entered here, the lemmas starting with these letters appear in a subwindow, and it is possible to scroll untill the search lemma is located; when the lemma is clicked, its part of speech, discriminator, and all orthographically neutral forms belonging to the lemma appear; if an orthographically neutral form is clicked, all the source forms it represents, appear in another subwindow, cf. Figure 1, 2, 4, and 5.

BDepending on the search interest of the user, a lemma, an orthographically neutral form or a source form is marked, and the information from the dictionary is subsequently transferred to a search window by clicking the relevant **Søg**-button. In the search window, several criteria may be combined by repeated consultation of the dictionary. By using the dictionary in this way, it is possible to construct innumerable searches. A search combining the plural form of the noun *rune* (En. *rune*) and the lemma *kaste vb* (En. *throw*) locates the ballads where someone tries to obtain love by means of magic. A search asking for the orthographically neutral form *hjertet* (definite form of *hjerte*, En. *heart*) occurring in the same verse as the adjective *kær* (En. *dear*), finds all occurrences of the phrase *haver i hjertet kær* (En. *loves dearly*). Figure 6 shows the combined search criteria. Figure 7 shows an extract from the results of the search.

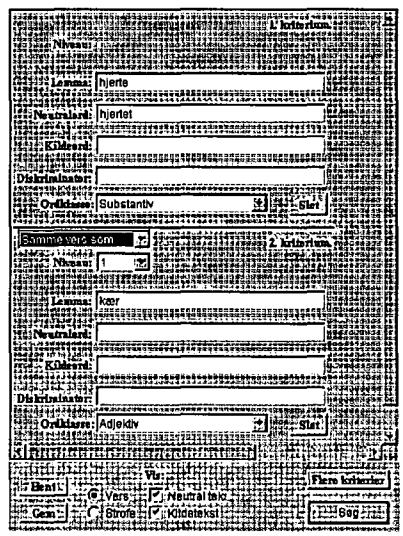
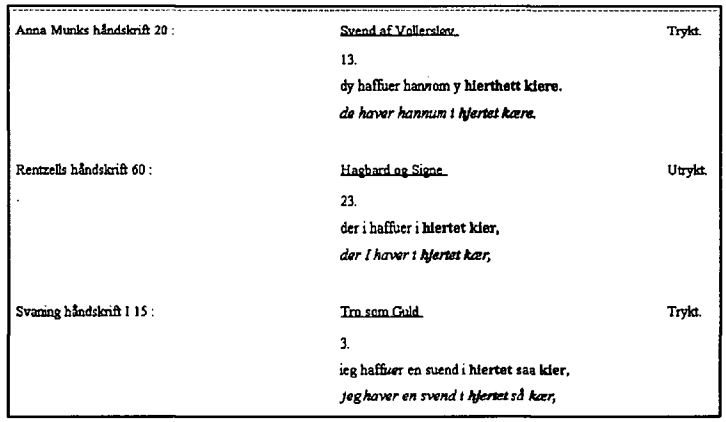


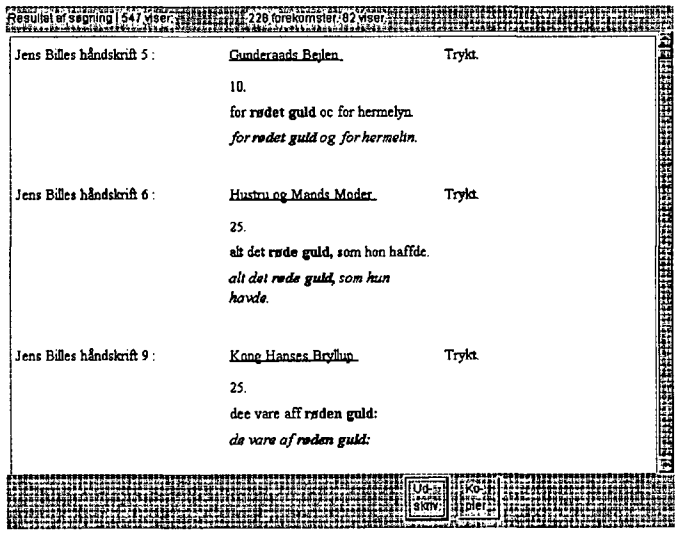
Figure 6: Two search criteria are combined by the operator *Samme vers som* (En. *in the same verse as*): *hjertet* (En. *the heart*) an orthographically neutral form of the lemma *hjerte* with the part of speech *Substantiv* (En. *noun*) combined with the lemma *kær* (En. *dear*) with the part of speech *Adjektiv* (En. *adjective*). At the bottom of the search window, it is specified that the occurrences found will be shown in source forms and in orthographically neutral forms, and that the context is the verse of the occurrence.



Anna Munks håndskrift 20 :	Svend af Vollerslev	Trykt.
	13.	
	dy haffuer hannom y **hierthett kiere.**	
	*de haver hannum i **hjertet kære.***	
Rentzells håndskrift 60 :	Hagbard og Signe	Utrykt.
	23.	
	der i haffuer i **hiertet kier,**	
	*der I haver i **hjertet kær,***	
Svaning håndskrift I 15 :	Tro som Guld	Trykt.
	3.	
	ieg haffuer en suend i **hiertet saa kier,**	
	*jeg haver en svend i **hjertet så kær,***	

Figure 7: Extract of results from the search in Figure 6. The words searched for are in bold face; for each occurrence the name of the manuscript (Da. *håndskrift*), the title of the ballad, and the number of the stanza are given. In the search results the title of the ballad is linked to the whole text, consequently, the broader context of the occurrence is easily available.

In the universe of the ballads, the possession of gold is a sign of social prominence, a search for the lemmas *guld* (En. *gold*) and *rød* (En. *read*) in the same verse yields 228 occurrences from 82 different ballads, an extract of the search results is shown in Figure 8.



Resultat af søgning | 547 viser; 228 forekomster; 82 viser.

Jens Billes håndskrift 5 :	Gunderaads Bejlen	Trykt.
	10.	
	for **rødet guld** oc for hermelyn.	
	*for **rødet guld** og for hermelin.*	
Jens Billes håndskrift 6 :	Hustru og Mands Moder	Trykt.
	25.	
	alt det **røde guld,** som hon haffde.	
	*alt det **røde guld,** som hun havde.*	
Jens Billes håndskrift 9 :	Kong Hanses Bryllup	Trykt.
	25.	
	dee vare aff **røden guld:**	
	*de vare af **røden guld:***	

Udskriv. Kopier

Figure 8: The results from a search on the lemma level show different forms of the adjective *rød* (En. *read*) *rødet*, *røde*, *røden* collocated with the noun *guld* (En. *gold*).

To follow the trace of gold, the dictionary is consulted; a look-up reveals that 20 compound words have *guld* as their first part. Another evidence of high social standing is the use of silk. The dictionary contains 15 compound words with *silke* as their first part. These observations fit in with the ideals of the owners of the ballad manuscripts, the great nobles of Denmark-Norway in the latter half of the 16th century spent a lot of effort on combining, dividing and fighting over manors and estates.

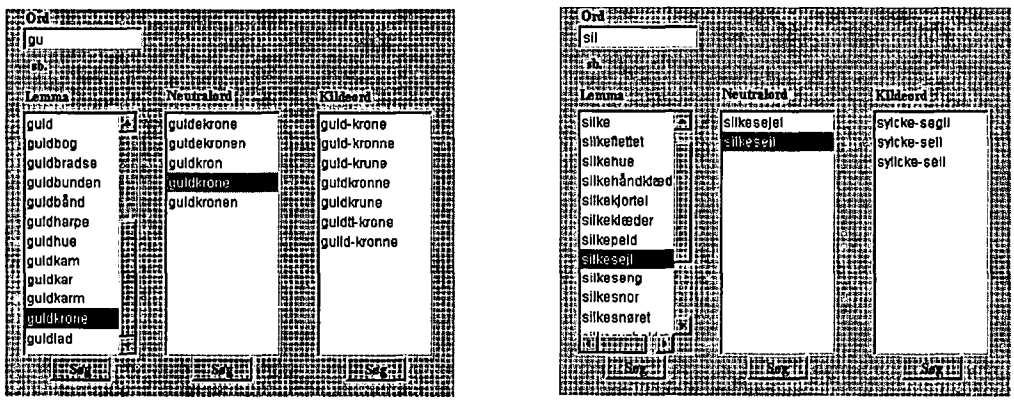


Figure 9: The dictionary shows how gold and silk figure prominently in the ballad universe as can be traced in the many compound words with *guld* and *silke* as their first part.

The Descriptive Levels as Search Tools

Depending on the user's search interests, one or several descriptive levels are involved in the specification of the search. A user interested in the material culture of the texts will concentrate on the lemma level and make searches based on the nouns of shipping, clothing etc. Gunner Lind [1999] has e.g. shown how the weapons play a central role in the ideal masculinity in the period by searching for occurrences of weapon nouns in the texts. A user interested in the oldfashioned adjective forms in *-en* and *-er*, will concentrate on the level of orthographically neutral forms. A user with special interest in the formulaic style of the ballads will combine these two levels, e.g. instances of a formula for arriving like *kom ridende i gård* 'entered on horseback' will be found by combining the orthographically neutral form *kom* of the lemma *komme* (En. *come*) with the lemmas *ride vb* and *gård sb* (En. *yard*). Syntacticians will make use of the orthographically neutral forms to search for grammatical constructions, e.g. the passive verbal forms *bads* and *bedes* from the entry *bede* (En. *ask*), cf. Figure 4. The source level forms are of special interest to philologists studying the characteristics of (parts of) single manuscripts.

The phrase *kom ridende i gård* indicates change of scene, the search combining the three main words reveals that the phrase has several variants: both *ridende, i*, and *gård* have variants on the orthographically neutral level - and of course numerous source level variants. The search further reveals that in the great majority of the cases, the person arriving on horseback is male.

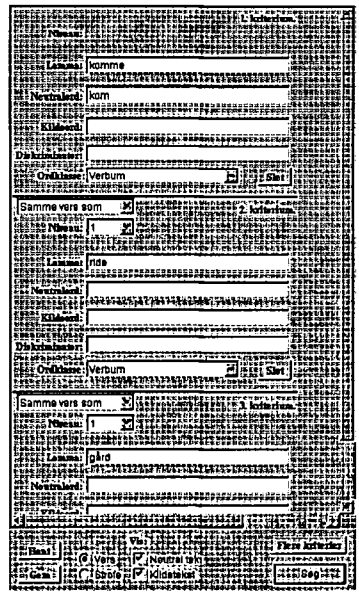


Figure 10: The combined search criteria for finding the occurrences of the phrase *kom ridende i gård*.

Research Examples

During the construction of the multi level corpus-base and the dictionary, scholars from several disciplines in the humanities seized the opportunity to make searches with the current dictionary in the current corpus-base and used the results in their respective investigations. As the methodology ensures that the current dictionary covers all the texts in the current corpus-base, and the current dictionary and base contain only data that have been carefully proofread, the results of the searches in the construction phase are completely reliable for the amount of texts encluded in the corpus-base. The only caveat for users in the construction phase is that the number of occurrences or lack of occurrences of a given phenomenon may change when more texts are included. The studies completed during the construction phase include historical pragmatics, the poetic formulae of the ballads, the ideal male of the 16th century and others.

The social variation in pronouns of adress

Most ballads have been recorded in several different versions; in the oldest tradition, 18 ballads and songs are recorded in five or more versions [Ruus 2000]. One of the characteristics of the ballads is the combination of narrative and dialogue, the dialogues between the main characters being fairly stable in the different versions. By searching for key words in greetings like *hel adj* and *velkommen adj* (En. *hail* and *welcome*), and forms of personal pronouns, it was possible to describe the modes of adress; a pattern based on social standing was revealed: persons of equal social status adress each other with the second person singular form of the personal pronoun, while older persons and persons of higher social status are adressed with the second person plural

form of the personal pronoun; correspondingly, *velkommen* is used when the persons are acquainted with each other, whereas *hel* is used in more formal situations [Ruus 1999].

Roses and lilies

In the poetic language of the ballads and songs, the maidens are often referred to as roses or lilies. These flower terms occur in a series of compound words, many of which are also used about young women. Vibeke A. Pedersen [2001] has described the formula system in one of the manuscripts by searching for key words in the formulae. Her studies show among others that both compounds like *liljekvist*, *liljevånd*, *rosenkvist* (En. *twig or branch of lily*, *of rose*), and *rosenblomme* (En. *rose flower*) are used for maidens. Other studies have shown the differences in the poetic language of the ballad genre as opposed to the lyrical songs of the manuscripts [Duncker 2000; Thomasen 2000].

Variation and stability in the lexical core

The inventory of lemmas in the dictionary and their frequence in the text corpus may form the basis for a diachronic study of the variation and stability in the vocabulary of Danish. A preliminary investigation comparing the frequent lemmas in the electronic dictionary of ballad and song texts with an inventory of core words from Modern Danish [Ruus 1998], has shown that the stable areas of the vocabulary comprise function words like conjunctions, pronouns and prepositions, and the vocabulary of semantic fields like colours and communication, while words designating persons and their social functions vary across time.

Conclusion

The corpus-based electronic dictionary of ballad and song texts from the 16th century is a priceless tool for all kinds of users, offering information about all the words in 547 ballads and songs - more than half of the oldest tradition - and giving access to all their occurrences. As the rich orthographical variance is captured in the dictionary, it is also possible to use the dictionary as a tool for searching in the 392 ballads and songs stored on the cd-rom in source level form only [Ruus 2001b].

The interested layman gets a unique key to the oldest ballad tradition. The orthographically neutral forms and the lemmas close to Modern Danish accompanied by explanatory glosses assist the amateur reader in his interpretation, and the source level text forms and the pictures from the manuscripts constitute a direct link to the ballad lovers of the 16th century.

For scholars of many disciplines, the corpus-based electronic dictionary opens innumerable avenues of study. Literary scholars may trace the themes and their poetic guise across manuscripts and ballad types. Historians may peruse the ideals and interests of the 16th century nobility, tracing important themes by searching with the dictionary. Philologists and historical linguists may compare the vocabulary and the phrases internally in each manuscript, and through the centuries, pursuing specific words and phrases in older and in younger texts.

The corpus-based dictionary approach described here is equally applicable to other collections of texts in languages with a lexicographical tradition, whether old or modern, spoken or written.

The multi level representation of the texts and the multi level information in the dictionary make it possible to keep all characteristic features of the authentic spoken or written texts, whether pronunciation variants, spelling variants, typographical errors or capitalization, linking the variants to the neutral forms suitable for searching.

The accumulating technique has the beneficial effect that the fully automatic processes carry an increasing proportion of the workload as more texts are processed [Duncker & Ruus to appear].

The careful combination of fully automatic processes, used when speed and repetition are the dominating factors, and of processes fully controlled by the lexicographer ensures that the markup of the texts and the dictionary based on the texts are as correct as humanly possible.

References

[Dansk Sprognævn 1986] Dansk Sprognævn 1986. *Retskrivningsordbogen*, København.

[Duncker 2000] Duncker, D., 2000. Orden i Viserne, in [Lundgreen-Nielsen & Ruus 2000].

[Duncker & Ruus 2000] Duncker, D. & H. Ruus 2000. Multi Level Text Representation in an LCB, in Jens Erik Mogensen et al. (eds.) *Symposium on Lexicography IX, Proceedings of the Ninth International Symposium on Lexicography April 23-25, 1998 at the University of Copenhagen*, Lexicographica, Series Mayor, Band 103, Max Niemeyer Verlag, Tübingen, p. 77-97.

[Duncker & Ruus to appear] Duncker, D. & H. Ruus to appear. Multi Level Text Markup. The Accumulative Approach.

[Lind 1999] Lind, G. 1999. Våbnenes tale. Våben, drab og krig i viser og virkelighed i Danmark 1536-1660, in [Lundgreen-Nielsen & Ruus 1999].

[Lundgreen-Nielsen & Ruus 1999] Lundgreen-Nielsen, F. & H. Ruus (eds.) 1999. *Svøbt i mår*, bind 1: Adelskultur og visebøger, C. A. Reitzel, København, 428 pp.

[Lundgreen-Nielsen & Ruus 2000] Lundgreen-Nielsen, F. & H. Ruus (eds.) 2000. *Svøbt i mår*, bind 2: Et spørgsmål om stil, C. A. Reitzel, København, 510 pp.

[Lundgreen-Nielsen & Ruus 2001] Lundgreen-Nielsen, F. & H. Ruus (eds.) 2001. *Svøbt i mår*, bind 3: Tæt på visetekste rne„ C. A. Reitzel, København, 516 pp. med indlagt cd-rom.

[Lundgreen-Nielsen & Ruus 2001] Lundgreen-Nielsen, F. & H. Ruus (eds.) 2002 [in print]: *Svøbt i mår*, bind 4: Lærdom og trolddom, C. A. Reitzel, København.

[Pedersen 2001] Pedersen, V. A. 2001. Formler i *Dronning Sophias visebog*, in [Lundgreen-Nielsen & Ruus 2001].

[Ruus 1996] Ruus, H. 1996. Das Forschungsprojekt *Dänische Balladenkultur 1550-1700* stellt sich vor, in O. Holzapfel et al. (eds.) *Jahrbuch für Volksliedforschung* 1996, Erich Schmidt Verlag, 109-111.

[Ruus 1998] Ruus, H. 1998. Viseord og Kerneord, Konstans og variation i det centrale ordforråd, in K. Kristensen (ed.) *Selskab for Nordisk Filologi, Årsberetning 1996-1997.*

[Ruus 1999] Ruus, H. 1999. Folkevisedansk - den ældste viseoverlevering på cd-rom, in P. Widell & M. Kunøe (eds.) *7. Møde om Udforskningen af Dansk Sprog*, Aarhus.

[Ruus 2000] Ruus, H. 2000. Visernes top-18. Populære viser i overleveringen før 1591, in [Lundgreen-Nielsen & Ruus 2000].

[Ruus 2001a] Ruus, H. 2001a. Elektronisk viseforskning, in [Lundgreen-Nielsen & Ruus 2001].

[Ruus 2001b] Ruus, Hanne 2001b: *Svøbt i mår*-cd-rommen, in [Lundgreen-Nielsen & Ruus 2001].

[Thomasen 2000] Thomasen, B. Brinkmann 2000. Det lyriske *sig* - om brugen af det refleksive pronomen i de lyriske viser 1530-1630, in [Lundgreen-Nielsen & Ruus 2000].