Knowledge for Everyman from the Renaissance to Modern Times

The Danish CLARIN project plans to deliver both a technical platform (an infrastructure with search and retrieval facilities) and platform content. The content will take the form of examples of many different and interesting types of resources: monolingual written corpora, aligned multilingual written corpora, monolingual spoken corpora, dictionaries, word nets, images, videos, etc.

One such resource for the use of the humanities research community is a monolingual written Danish corpus of approximately 250,000 words composed of extracts from non-literary texts for everyman's use from the period 1500 to 1750. The texts will be extracted from rare books only obtainable from The Royal Library in Copenhagen, and they will cover subjects such as ethics and moral issues, geography and topography, history, housekeeping and cooking, medical science, mathematics and astrology, natural sciences, pedagogics, etc. Through the CLARIN platform the texts will be made available electronically, marked up, and with a dictionary as a lexical key to the corpus.

How to look up a word with orthographic variation?

The texts are written by different authors and over a period of time when orthographical rules for written Danish had not been stabilized, so there will be a high level of orthographic variation in the corpus. Examples of such variation can be seen in the Danish word currently spelled *sygdom* (en: illness), which can have the following spellings in older texts: *sigdom, siugdom, siugedom, sygdom, sygdomme, sygdommer, siuge, syge, syuge.* It is necessary to neutralize such spelling variations in order for the researcher to be able to search for and find instances of certain words or phrases in the texts.

The DUDS research group at the Department of Scandinavian Studies and Linguistics at the University of Copenhagen, which is in charge of producing the corpus, has developed a neutralization method and a mark-up system suitable for texts with orthographical variation. They have developed the method and successfully implemented it on the Danish ballad manuscripts of the 16th century thus making the complete textual tradition before 1591 available electronically with a dictionary. [1].

[illustration: An opening of the ballad manuscript called *Hjertebogen*,
en: the heart book <<http://www2.kb.dk/elib/mss/skatte/mss/thott_1510.htm>>
With permission from The Royal Library, Denmark]

The method is called multilevel text, MLT, and it consists of providing three linked levels of markup on each word in the corpus: source level, neutral level, and lemma level. The source level is the original word form as written in the text or manuscript; the neutral level represents a neutral word form close to modern Danish spelling, and the lemma level gives the lemma form of the source word with the associated part-of-speech. To illustrate the richness and complexity of orthographic variation we present an extract of a search result for the neutral form *hjertet* (en: the heart) from the MLT marked-up ballad corpus (only source level forms shown):

Then første hand var y	**hiertiidtt**	gladtt
dj iegh er ham aff	**hierttett**	huldtt,
och saa den frue, du haffuer y	**hiertet**	kier.
the hellede wore y	**hierttit**	trøst,
och elsk hanom aff	**herttiitt**	
det oden i	**hierted**	vende:
da maan ieg ham aff	**hiertted**	gaa,
y maa vell sige aff	**hiertet**	frÿ:
hun hagde stoer soriig y	**hierthet**	sin,
i haffuer meg i	**hierthed**	saa kier.
de haffde huer-andenn udi	**hiertid**	saa kier,
tro myg, yeg dyg aff	**hyerthett**	for nogin mand well vntt..

Research Examples

The Knowledge for Everyman corpus is still being built and no research has yet been based on findings in the corpus. The possible research themes, however, are many, for example the following ones suggested by the DUDS researchers: *Perceptions of and attitudes to health and illness in the period 1500-1750* (based on medical books and cook books, using search terms for e.g. fresh, health, ill, illness, medical.), *The use of medical herbs* (based on medical books and cook books, using the names of the herbs as search terms), *Religion in everyday life* (based on catechism, prayer books, ethics, and with search terms for e.g. christian, pray, prayer, enemy), and *Knowledge about the world* (based on descriptions of exotic countries, pamphlets about sensational incidents and the like, and using the geographical names as search terms).

The corpus of ballads from before 1591 with the ballad dictionary is already a rich resource of information for scholars from different disciplines in the humanities. Many scholars have made searches in the corpus-base and used the results in their research covering themes such as *Formulae*, e.g. *The poetic formulae of the ballads* which studied the flower terms involving roses and lilies used to refer to young maidens. Other themes were *Weapon* studied by a historian (what is said about weapons in the ballads, how does this correspond to weapons, killing and war in the society), *Pragmatics*, for instance *The social variation in pronouns of address* (studied through key words in greetings combined with forms of personal pronouns), *Orthography*, *Stylistics*, *Genre definition*. Precise references to the research mentioned here can be found in [1].

Conclusion

Today the corpus of ballads and its dictionary is available on CD-ROM together with volume 3 of [2]. The 18 most popular renaissance ballads and their textual variants are available with neutral word forms at
http://duds.nordisk.ku.dk/tekstresurser/aeldste_danske_viseoverlevering/visernes_top-18/, and the

remainder of the ballad corpus is being prepared for publication on the Internet. The corpus-to-be of knowledge for everyman is not yet available electronically.
The CLARIN infrastructure aims to become the common vehicle for taking traditional texts for humanistic research into the electronic future where they can be made available to other researchers, not merely in their original source form, but also with mark-up and other enhancements produced by research colleagues and with tools for further enhancement.

References

[1] Ruus, Hanne: A Corpus-based Electronic Dictionary for (Re)search, in EURALEX 2002 Proceedings, pages 175-185.
[2] Lundgreen-Nielsen & Ruus, Hanne (Ed.): Svøbt i mår, bind 1-4, København 1999-2001.

Thanks to Hanne Ruus and Dorte Duncker of the DUDS group for inspiring conversations and useful comments.