Corpus-Based Word Information Systems for the German Language on the Internet:
Aspects of Usability and Applicability in the Focus of Scholarly Needs[1]

Thomas Schares

thomas.schares@uni-hamburg.de
Goethe-Wörterbuch [Goethe-Dictionary],
Arbeitsstelle Hamburg, Universität Hamburg, IfG II, Germany

Abstract
In this paper I will describe usage possibilities of two major lexicological resources for the German language on the internet. The two databases in my focus will be the elexiko-project by the IdS (=Institute for the German Language), Mannheim, and the DWDS-project by the Berlin-Brandenburg Academy of Letters & Sciences. Both items are corpus-based and by self-definition claim to be comprehensive word information systems of German.

Eventually, I would like to put the user perspective on a scholarly level and I will try to pursue the view of a professional user – in this case a linguist and/or German philologist. I will apply two fairly simple and straightforward search-queries derived from linguistic research to both resources. Thereby, first, I will try to show what kind of material and information can be obtained from both databases in their present state. Second, an attempt to summarize the specifics to be fulfilled in order to meet the needs of a professional in like databases will be made and a conclusion in the form of some statements will be given of what can be expected from web-based word-information-systems, what can – and what should – be provided to embrace the needs of the professional user.

Hitherto, aspects of the (professional) usability of web-based lexicographical/lexicological databases have not been examined widely; I will stress on the importance of the anticipation of user-needs and the application of user typology for the design and concept of such projects.

1 Introduction and aims
In May 2007 the relaunch of a major online project dedicated to the documentation of the German language was celebrated with a two-day colloquium (cf. Michaelis [to appear]) in Mannheim/Germany. It involved many fruitful discussions on various aspects of online-lexicography and its aims and duties, especially the necessity of user-oriented approaches has been emphasized repeatedly. Consequently I would like to pick up this notion and carry out an experiment on the usability of two German word-information-systems from a scholarly point of view. By this I would like to test the possible application of online lexicographical

---

[1] I would like to express my gratitude to Christiane Schlaps (Hamburg/Germany) and Frank Michaelis (Göttingen/Germany) for substantial help and inspirational support with this paper. All actions carried out in order to gather the material from the internet on behalf of this paper in the second half of 2007.

resources in linguistic research, i.e. in real scholarly life. Hitherto, aspects of the (professional) usability of web-based lexicographical/lexicological databases and of electronic dictionaries in general have not been examined widely,[2] although the need for a user-oriented exploration of the applicability of like systems is generally stressed. I have chosen two major resources for the German language on the internet, elexiko and the DWDS, to carry out my experiment. Both are prominent and institutional projects[3] and both are dedicated to the lexicographical documentation of the German language; they are representatives of the „changeover to digital lexicography" (Kramer 2006, 51). They are comparable since both are corpus-based and by self-definition claim to be comprehensive word-information-systems of (20th century and contemporary) German, and both also emphasize to be projects-in-progress.

I lack the space here to give a detailed description of the partly complex genesis and history of both projects, but a few words on them will suffice[4], since my aim is not a description of the full range of features offered in both databases but their suitability to support a linguist in exploring predefined problems.

elexiko is a „lexikografisch-lexikologisches Projekt" 'lexicographical-lexicological project' (Klosa & Steffens 2007, 443): It consists of a dictionary with ca. 300.000 entries, presently ca. 700 entries are fully crafted (edited) dictionary entries. It also offers a number of features that go well beyond the scope of a dictionary, some of which will be introduced in the following.[5] The DWDS-project aims at creating a corpus-based (new) electronic dictionary derived from a retro-digitized (print) dictionary[6], access to the underlying corpus as well as to word information going beyond the traditional dictionary entry contents are also promised. This additional information on words comprises, e.g., issues of collocation in elexiko („lexikalische Mitspieler" 'lexical mates') and in the DWDS („Wortverbindungen" 'collocations'). This in-a-nutshell description of both projects indicates at the hybridity of such word-information-systems: They – typologically – are somewhere in between a dictionary and a corpus-search tool. In my conclusion I will come back to this notion.

2 Test queries
2a *ieren-verbs
In order to test the abilites and qualities of both word-information systems from a linguist's point of view two test queries will be applied to them. The first test query is relatively

---

[2] Cf. Haß 2005, 4; but cf. Jucker 1994 for an early instrumentalization of the electronic OED[=Oxford English Dictionary] in vocabulary studies, cf. also Lemnitzer (2001).
[3] The elexiko-project [elexiko = <u>el</u>ektronisches, <u>lexi</u>kalisch-lexikologisches <u>kor</u>pusbasiertes Informationssystem 'electronic, lexical-lexicological corpus-based information system'] is fostered by the IDS [= Institute for the German language]; the DWDS [= Digital Dictionary of the German language of the 20th century] is maintained by the BBAW [= Berlin-Brandenburg Academy of Sciences]; URLs: (institutions:) http://www.ids-mannheim.de and http://www.bbaw.de; (projects:) http://www.elexiko.de and http://www.dwds.de; there exists another corpus-based word-information-system at Leipzig University (wortschatz.uni-leipzig.de) but due to the level of expenditure for this paper and the limited functionality of that resource as well as the lack of available information there I have not included it here.
[4] Cf. Geyken & Klein 2001, 263-270 for the DWDS and Haß 2005, 1-35 for elexiko.
[5] For a full description of the – relaunched – elexiko cf. Michaelis [to appear].
[6] The WDG [=Dictionary of Contemporary German] compiled in the former German Democratic Republic between 1952 and 1977, for information on DWDS-corpus cf. Scherer (2006), 76ff.

straightforward and not complex. It is taken from actual research carried out and in spite of its simplicity it provides an interesting example for the applicability of searches carried out on electronic texts: The German infinitive of verbs serves as a base form (lemmatized form) in dictionaries and metalingual texts. There exists a number of sub-groups to the standard infinitive-form on -en. One very interesting sub-group of these is the -ieren-infinitive, which historically has been derived from a french verbal-morpheme in the middle-ages.[7] This group of verbs is still productive in word-formation and therefore it should be highly interesting for a linguist to obtain all occurences of this derivation from a corpus in order to estimate on the size and the productivity of this group of verbs and this morpheme.

2a1 *ieren – elexiko

The portal-page to the elexiko-project gives quick access to the database by a single search-box. The search string „*ieren" yields an immediate result; server response, however, takes a while (in several tests it took an average time of 10 to 15 seconds for response). The result to the search query is a list of words ending on -ieren. The list comprises only verbs, as desired and it is sorted alphabetically. This indicates that the textdata in the database is lemmatized, i.e. only uninflected word-forms are in the database, inflected word forms like *abkommandiert, studierte* as they appear in the corpus (tokens) are represented by their infinitive in the case of verbs (types). This has two immediate advantages for a relatively straightforward search query as the one in this example: First, it is not necessary to carry out a number of search queries for different possible inflected verb-forms, since they are all searchable by their infinitive which always ends on -ieren. Second, no other non-verbal word forms like *Offizieren* are found accidentally (cf. 2a2). It is not necessary to sort out such possible collateral findings from the results-list subsequently. The number of results found in the database is given at the end of the list. In this case: 2276. The results-page also provides a new-query-button which leads back to the portal-page and a save-results-button, which, unfortunately, doesn't always work properly.[8]

It is also possible with not much more effort to determine the occurence frequency (in types!) of two sub-types of -ieren-verbs, the two derivational types of *-isieren* (*elektrisieren* 'to electrify') and *-ifizieren* (*mumifizieren* 'to mummify'). The results obtained are in summarized form:

-ieren	2276	*-ifizieren*	46
		-isieren	372

It is clearly visible that one sub-type is much more productive than the other and also that the base-type is by far dominating. This is a useful first result, and it is reliable since the figures are covered by corpus evidence. An in-depth analysis of the material obtained from Elexico, however, is not as easily attainable. The results list merely yields the isolated (lemmatized) word-forms, only types, no tokens. Within the elexiko-platform the user has no possibility to convert the results-list into a KWIC (key-word-in-context)-index. To see how a word from the results list works in context, the user has to open the entry to the word (by a click on the word in the list). Here, examples of usage can not be viewed immediately, two more clicks are necessary, before examples of usage can be viewed. The button „print version" produces a

---

[7] Fleischer & Barz 1995, 311f; Eichinger 2000, 155; I will not pursue a discussion here whether the morphological type of these verbs is derivational or conversional.

[8] Tried with Opera 9.1, IE 6.0 and Firefox 2.0 with error message or zero activity, these browser-versions also for all browsing activities carried out for this experiment.

page with all components of the dictionary entry visible: every information class available in the full entry of elexiko is listed here. In a fully crafted entry like AKZEPTIEREN this is a bulk of information, an unedited entry though does not contain a lot of lexicographical information (e.g. AKZENTUIEREN). The usage examples given are lexicographically motivated examples and serve to illustrate the meaning. For a deeper analysis of usage in context (like in a KWIC-index) these are not sufficient. Moreover, comparing the usage of different verbs of the *-ieren*-class becomes time-consuming since a lot of different entries would have to be accessed separately in order to collect the usage examples. Furthermore, these are only available in the fully crafted entries (presently ca. 700 out of 300.000). A reversed headword-list is also available, it is a little long-winded to operate though: the matches for *ieren can not be opened in one single list but only in portions; in addition, the reversed headword list portions can not be exported as textfiles.

Yet, a user familiar with the history and genesis of the elexiko-project (which, at any rate, can not be assumed of the average user) knows that the corpora forming (part of) its basis are accessible by another door[9] which possibly may serve as a short-cut: one of the most important historical merits of the IDS, the mother-organization to elexiko, is the provision of (electronic) text-corpora to the German language. These are accessible and searchable via COSMAS II (Corpus Search, Management and Analysis System, also developed at the IDS[10]). To have access to this resource, the user has to register beforehand. The processing of the query in the COSMAS-database was comparable to the experiences in the DWDS (see section [2a2] below), therefore I will not go into details here. It is noteworthy, however, that operating the complex COSMAS search engine is a complicated task.

elexiko also offers the feature of a co-occurrence analysis. By clicking a button („CCDB") in the entry display field it can be applied. This „CCDB"-feature has also an entry in the help glossary but I was not able to open this entry, not even by trying with different browsers. The example *akzeptieren* shows the outline of this feature. In the co-occurrence list, the words that go together with the target word are listed along with figures and percentages. The list presented needs explanation, especially the figures and percentages given, but the glossary entry in the help section unfortunately is not available, as mentioned. The feature furthermore does not work properly, for some random entries (e.g. *harangieren, karresieren, präkonisieren*) it produced an error screen.

2a2 *ieren – DWDS

That the user has to register before being allowed to use a service on the internet seems to be common practice in the scholarly community by now, even though it is does not become immediately apparent, why a user of COSMAS (see above) or the DWDS should register, since neither charging nor (overt) logging procedures are employed. However, the DWDS can also be used without registration, but an unregistered user will only have access to a limited amount of results. In the *-ieren*-example 6866 compared to 81800 (in the following screen the numbers are 6866 vs. 74934, an inconsistency I cannot explain). The only difference which has become obvious to me between the logged-in- and the logged-out-status in the DWDS was that while being logged in, the database produced more error-pages. Limitations of access were still valid.

---

[9] Which is not mentioned on the elexiko-homepage, although; – why this?
[10] http://www.ids-mannheim.de/cosmas2/, cf. Scherer 2006, 80-85.

On the start page of the DWDS it becomes clear that here a dictionary and corpora are provided. The user can access both via the searchbox. Truncation also works here: the search-string „*ieren" is processed. In the results list (the only results found are in the corpora, no results in the dictionary section) one can see, that the word forms in this database are not lemmatized: Thus, also non-verbal word forms such as the nominal plurals *Tieren* ('animals'), *Infanterieoffizieren* ('infantry commissioned officers'), *Lightbieren* ('light beers') or *Wertpapieren* ('securities') are found. A search for *akzeptieren* ('to accept'), however, also yields inflected forms of the verb, so access to lemmatized forms seems to depend on the structure of the search-string. Lemmatization and search option seem to collide when truncation is employed (again, a mixup of types and tokens); this feature runs more smoothly in elexiko (but there no corpus access is offered yet). Another problem in this context is that inflected word-forms were not found in the truncated serach string „*ieren"; only the search for a definite word form such as „akzeptieren" produced inflected verbal forms in the results list. Therefore the results obtained here cannot be processed directly in linguistic research on -*ieren*-verbs. It is possible to export the results as a KWIC-index. In this index the word forms are given in the syntactic context they appear in. References to the source-texts can be added to the KWIC-indices. The number of results in this export facility is limited to 500, although. This exported index can serve the linguist as a list of occurences, it can hardly provide a stand-alone-basis for actual linguistic research. When searching single word forms without truncation, such KWIC-indices from the DWDS can provide valuable and well-documented data for linguistic research.

The DWDS-page offers another interesting feature: Under „Wortverlauf" ('chronological appearance of word/searchstring') a graphical analysis of the chronological frequency of corpus appearance can be generated. Applied to the **ieren*-query a considerable rise of word-forms ending on -*ieren* in the second half of the twentieth century is detectable. This is also ascertainable for „akzeptieren". Yet, this statistical device can indicate a very rough tendency only, since it is not clear how many word forms which I didn't want to be included into my search query were processed anyway (such as the numerous nominals appearing with the query „*ieren"). Type-token-relations are also not taken into account in such a rough analysis. The approach contained in this feature is undebatedly promising, but needs of linguistic methodology have to be obeyed before it can be seriously employed in linguistic analyses.

2b ins/in das
The second linguistic problem I would like to examine by employing both internet-resources in the focus of this paper is a question of word collocation. The combination of the preposition *in* followed by a definite neuter article *das* is commonly clipped to *ins*.[11] By observing the occurences in different sentences and contexts I would like to find out whether this is a free variant or if it has to obey certain distributional rules. Some „Funktions-verbgefüge"[12] ('functional verb constructions') and collocational and idiomatic usages like *ins Bett gehen* ('to go to bed'), *ins Kraut schießen* ('to increase rapidly'), *ins Gewicht fallen* ('to matter'), *zu tief ins Glas geschaut haben* ('to have drunk too much'), *ins Gras beißen* ('to die', derogatory) only work with the contracted variant, the uncontracted variant (**in das Gras beißen*) is considered ungrammatical by native speakers. I would like to find out with the help of a corpus search, if it contains variants of usage where the uncontracted variant is preferred. Maybe it is also possible to obtain evidence for the assumption that the use of the contracted

---

[11] In German there exists a number of such contractions like *ins < in das, ans < an das, ums < um das, im < in dem, am < an dem* etc.

[12] For a definition of „Funktionsverbgefüge", a peculiarity of German, cf. Winhart 2002, 5ff.

form has increased since the second half of the twentieth century. In sample query (3a) the occurence of derivates on -*ieren* has been in the focus (types). Concerning this query, the linguist will pay more attention of the examples of usage (tokens) in order to find out if it is true that certain usage contexts demand or at least prefer a contracted or uncontracted form. Therefore it will be of major importance to have access to usage examples.

2b1 ins/in das – elexiko
The query for the word *ins* leads to an hitherto unedited entry in the elexiko-dictionary. The lexicographical question arises here, if this word really should be lemmatized and form an own entry in a dictionary, since it seemingly is a phrase of two words which are contracted sometimes. However, no information on the word/the contraction can be obtained here, not even the insecure word-status of the expression is recorded here. Since I am especially interested in usage examples for the expression, I have tried to follow the link to the collocation listings („CCDB"). The link leads to an error prompt (It works although from other entries, e.g. „deutsch" and several of the *ieren-entries, cf. 2a1). The search for the pattern *in das* has no hits. Here, difficulties concerning the provisional status of the headword list become apparent: The contracted form *ins* has been included, but not its underlying base-form, the uncontracted form *in das*, – probably because it is not considered a word. In this case *ins* should not be considered a word and a proper dictionary entry, either. The headword-list is not consistent in this respect, since e.g. *ans* (contracted from *an das*) has not been included.[13] – To sum up: the dictionary-nature of elexiko becomes apparent here. Multi-word-expressions are not searchable. Consequently, elexiko is of little help for my exploration of *ins/in das*.

2b2 ins/in das – DWDS
A search for *in das* on the start page obtains no result. To get results anyway, the user has to switch to the corpus-search-page. Here the report „no results" is repeated. But in a line not very prominently visible the user is questioned „or did you mean 'in das'". The search phrase visible here is hyperlinked and a click on it opens the results list. The problem with these results is that the article *das* is obviously coded as definite article, regardless of the genus. Therefore *der, die*, *den, dem, dessen* are also found when searching *das*. The help-page[14] gives a hint to solve this problem: The search string can be encoded in order to search for the exact word forms without further filtering. The search string „@in @das" produces the desired result, morphological coherences are ignored and only the word combination *in das* is found.

The search for the contracted form *ins* is less difficult. Indices fo both forms of occurence can be obtained from the DWDS-corpus. Due to the limitation of hits for unregistered users full indices covering the whole corpus are not obtainable. An attempt to obtain a full index as logged-in user failed – for no apparent reason the system produced an error screen for any multi-word-phrase.

---

[13] E.g. the inclusion of the words *Flash, Inro*, *Macromedia* and also a number of proper nouns like *Ria, Marcia, Wisteria*, also found in the elexiko-headword-list, seems disputable and the examples support the assumption that the elexiko-headword-list needs revision, but here is not the place to discuss this.
[14] The help-page on the DWDS-page is to find once the user has opened the corpus-module; the truncation symbols that can be employed are introduced there, cf. also (3b).

2b3 Linguistic Conclusions on the Basis of Evidence Acquired from elexiko and the DWDS
As outlined, elexiko was not helpful for gaining material on the the problem presented. This is partly due to the nature of this problem being established somewhere in between word and phrase: *ins – in das.* And phrases are traditionally hard to find in dictionary headword-lists. But it has been demonstrated in (2a) that elexiko is not exclusively a dictionary but a word-information-system suitable for queries that involve a corpus as well. Word formation as being in the focus in (2a) could very well be analysed by applying headword lists (types) from dictionaries, in the case of suffixes as *-ieren* in reverse sorting sequence. But it was also important to have access to usage examples in (2a) in order to assess some more obscure occurences, i.e. some noun plurals. The availability of usage examples (tokens) certainly is essential for questions like (2b). Any linguistic statement on the distribution of contracted/uncontracted *ins* has to be based on evidence in a usage example. The occurence alone (like in a headword list) is useless for application examples like these.

The following conclusions based on the KWIC-indices from DWDS can be made: In idiomatic and collocational expressions like *ins Gras beißen* and also in more general collocations like *ins Bett gehen* the usage of the contracted form is fixed and the usage of the uncontracted form is considered ungrammatical. In deictic usages of *das*, the uncontracted form is preferred but not canonical. The uncontracted form is canonical in correlative usage („... ist ein Gründerzentrum entstanden, *in das* junge Wissenschaftler wechseln können, ...“ from DWDS-corpus). In all other cases, which could possibly be further categorized, no more overt distributional rules could be found. In total the contracted form is more frequent than the uncontracted form. Observations concerning the chronological rise of the contracted form could not be derived from the material gained during the experiment.[15]

3 Results and Conclusion
Two experiments carried out on the word-information-systems in focus have tested their usefulness for actual linguistic research:
For experiment (2a) the alphabetic headword list produced by elexiko was most helpful for the linguist in order to gain some substantial insights on the topic, i.e. the productivity of the derivational morpheme *-ieren*. In the DWDS the truncated search-string **ieren* unfortunately produced no results in the dictionary module but only in the corpus-module. Additionally, here inflected word-forms were not taken into account. Thus, the KWIC-index from the corpus module was interesting but of no substantial support for this query; the frequency anlysis also produced very rough data (due to the missing inflected forms as well as erroneously included nominal forms), which is interesting but not suitable for scholarly interpretation.

For experiment (2b) vice versa the KWIC-index produced by the DWDS-corpus-module was most useful, it took some effort although to obtain a correct index for the multi-word-entry *in das*. In this second experiment the provisional stage of elexiko became apparent, the dictionary entry to *ins* was not edited yet, and the search for the multi-word-phrase *in das* produced no results. The inchoate nature of the DWDS also got apparent when the system produced no results while I was logged in as a user.

---

[15] Please note well that I have ignored any literature possibly existing on the topic; the purpose of this paper is to present ways of application of the two word-information-systems to scholarly research.

In the following I would like to remark on some more topics only loosely attached to my experiment but of substantial impact on the (professional) usage of both databases and the application of user-studies, and then I would like to round off this paper with a general statement on the concept of online word-information-systems.

3a Documentation and Corpora
A detailed documentation and description of elexiko is only available in printed form, the information given on the homepage is scarce (and in some instances hard to find), the user interested in background even as elementary as corpus-composition has to turn to printed information for reference (Haß 2005 – costs: € 98 – and a publication, moreover, with information partly obsolete meanwhile, cf. Schlaps 2006). Information essential to the scholar like the composition of the corpus should necessarily be available online; e.g. the printed volume (or excerpts) could easily be made available as a pdf-document (but I shall not go into open-access-policy matters here). The diligent scholar is reluctant to use data obtained from a black box. The documentation of the DWDS is not much different, but at least some of the publications to the project are available as pdf-documents.

A closer look on the corpora employed also seems necessary: On the elexiko-homepage the user does not learn very much about the underlying corpus: it consists of 1,3 billion tokens and contains newspaper and journal texts. To gather more information on the corpus the printed volume on the project (Haß 2005, 66) has to be consulted: The corpus contains tokens obtained almost exclusively from newspapers from the second half of the 20th century; due to its unbalanced compostition is can be considered the project's „gravest problem" (Schlaps 2006, 312). In the printed volume to elexiko (Haß 2005) the user also learns that external access to the corpus is not yet possible because of legal questions to be answered beforehand. The DWDS-corpus composition is described on the homepage and it seems balanced; it claims to be „das erste zeitlich und nach Textsorten ausgewogene Textcorpus der deutschen Sprache des 20. Jahrhunderts " ('the first text corpus of 20th century German balanced according to chronology and text types') and it contains one hundred millions tokens.[16] Concerning the access to the corpora's contents: from a scholarly point of view it also seems highly desirable that the access to corpora is not limited, restrictions lead to sketchy results that can not be utilized when employing strict linguistic methodology.

3b Retrieval
The range of search options offered in the DWDS is an interesting example for the application of intricate tagsets in corpus-search. The tagset applied in the DWDS corpora is the STTS-tagset developed at Stuttgart University.[17] A fairy complex query syntax is involved with this; the user has to familiarize him- or herself with this before a search query can be applied correctly (the problems involved are discussed in 2b2). Ad hoc results will not always be procurable for the quick passer-by user: e.g. for a plain fulltext-search a truncation symbol (@) has to be placed in front of (each part of) the search string (cf. 2b2). By this marker a

---

[16] http://www.dwds.de/textbasis; it is interesting to note, however, that e.g. Stanulewicz (2007)reports to have applied a Polish corpus by the Polish scientific publishing house PWN (korpus.pwn.pl) which seems to be remarkable less for its size but more for its representativity concerning "text type, period covered, style, region, and age and gender of of the authors" (88) – hence a comparison to corpora of other languages concerning size and composition could possibly support a justified evaluation; the DWDS corpus claims to be comparable to the British National Corpus [BNC] in size and content.
[17] Cf. http://www.sfs.uni-tuebingen.de/Elwis/stts/stts.html; guidelines available there.

specialized search (grammatical/morphological) becomes inverted into an ordinary search (fulltext); that means the normal variant in this system is the special (marked) variant. The user has to be familiar with peculiarities like this before being able to work effectively with the search apparatus in the DWDS. The outcomes in the results are not always logical, e.g. I did not fully comprehend the in-/exclusion of inflected word-forms in the results to different search queries.

The help-file to the search options in elexiko is difficult to find. It cannot be found in the glossary but is only accessible via a separate link on the „Benutzungshinweise"-page. A direct link to the search-help at least from the expert-search-mask and from the search-box on the start-page would be helpful. Truncation with an asterisk as well as with a question mark for a single letter is possible in elexiko. Some more search options are offered in the expert search but do not work very well yet because of the lack of searchable material since the dictionary is not yet completed.

3c The Gulf of Execution – Users
It is comprehensible that such important projects on language documentation and description should be designed to appeal to a large scale of users from non-professional to professional. As a result of the experiment presented above two statements regarding this can be made:

(a) Concerning the technical handling of the databases every user should be considered non-professional (especially from a viewpoint of computer-science or computational-linguistics who should not be mixed up with linguistics), in order to keep the level of frustration low. The design and content presentation should be comprehensive and intuitively operable. Features like the co-occurrence analysis in elexiko and the formal search-language in the DWDS are innovative features but hard to operate and comprehend even for the seemingly „professional" user (for a linguist is not necessarily a computer-scientist, or even familiar with PC-applications)[18], consequently special attention has to be paid to operability aspects – the gulf of execution – when designing expert search features, which sounds trivial but, as practical application shows, is not.

(b) Users without scholarly background should be welcome, but the focus in projects like those introduced here should be on scholarly interests. Scholars, linguists are interested in more advanced usage features like corpus access, co-occurrence and frequency analysis as offered in such projects; the accidental passer-by will hardly play for long with any of these features and to what purpose? Therefore simplified (and misinterpretable) metalanguage like in elexiko (e.g. "Lesart" for "meaning", a term denoting an alternative reading in critical editions usually) or linguistically imprecise annotations like in the DWDS (e.g. an inflected word form to the linguist is a grammatical, not a lexicological phenomenon) are not useful to the scholarly user but rather have a confusing effect. Exact linguistic terminology can be explained for the non-professional by easily accessible and flexible help features.

---

[18] Cf. also Schlaps 2006, 312, on elexiko: „... die Wortartikelansicht [orientiert] sich eher an den Bedürfnissen von Laien, die Recherchefunktion eher an linguistischen Fragestellungen" - 'The dictionary entry presentation is designed for the needs of the non-professional user, the retrieval mask is rather designed for linguistic research.'

In the course of this study it has become clear that the notion of a professional versus a non-professional user is fairly global. The tacit equation of professional user and linguist certainly has to be refined in more elaborate studies. Also, computer scientists and computer linguists have different professional levels. One very important type of non-professional user is the L2-learner, who certainly has distinct user needs. The whole issue of the dictionary-user-question is summarized adequately in the following statement: (a) Who uses a dictionary (b) how (c) when (d) how long (e) where (f) why (g) to achieve which goal (h) with how much success? (Wiegand 1987, translated from German and slightly changed). Dictionary makers *ex officio* are obliged to answer this question at least in part. (cf. 3f)

3d Some More Desiderata

Some more aspects and features seem desirable from a professional user's point of view. Among these are trivia like better server performance, elimination of bugs, and functional user-interfaces, the above inspection has shown that a closer look at such trivia from the developers is well worth it. Some expert desiderata are: exportable, corpus-dervied KWIC-indices with source information (quotability, documentation), and separately searchable (reversed) headword-lists (augmentable by e.g. word-form-indices, which would be helpful for queries like 2b), all in all a range of word-lists that properly take into account type-token-relations (cf. Hunston 2006, 235). Also, more transparency in information policy (project-documentation, help, guided tour/systematic introduction[19] etc.) would be highly desirable. It is not sufficient for online-projects to offer printed documentation only on vital aspects of the project. Finally, and this is also one of the trivia but highly desirable, the completeness of the dictionaries: It is equally frustrating for professional as for non-professional users to run into unedited dictionary-entries repeatedly, and it is well bequeathed knowledge from the print sector that completed dictionaries have a higher status of acceptance than unfinished ones. But concerning this, complete availability does not seem to be a matter of years than rather decades.[20]

3e Convergence Into Hybrid Systems

elexiko and the DWDS aim at enriching the concept of the traditional dicitionary by a number of additional features, which typologically do not belong to dictionaries and to some extend even contradict the concept of a dictionary: e.g. in a dictionary selected examples of usage are given in order to present typical examples, the process of selection is fundamental in dictionary-making; in a corpus search, in comparison, every (intra-coporeal) occurence of a search phrase without preselection is of interest. These two notions are not naturally compatible in a hybrid project. A corpus search KWIC-index can necessarily not be (an immediate) part of a dictionary entry, also the direct linking of a dictionary usage example to the full text is controversial among lexicographers. Therefore, and as a consequence of the findings of this paper, it is recommendable to distinguish the dictionary from the corpus component in such word-information-systems, and in the process of convergence that becomes apparent in the two projects visited here, a methodological awareness of this protean nature of this new enriched types of dictionaries should be advocated. Therefore I would suggest a clear division between dictionary contents and corpus access – which to some extend is present in the DWDS, whereas elexiko aims at a slightly different direction. In the present state elexiko is primarily a dictionary, and the DWDS is primarily a corpus. The following figure illustrates the aforesaid:

---

[19] In elexiko, the implementation of a guided tour is announced although: (29.08.2007)
http://hypermedia.ids-mannheim.de/elexiko/Portal/Aufgaben.html.
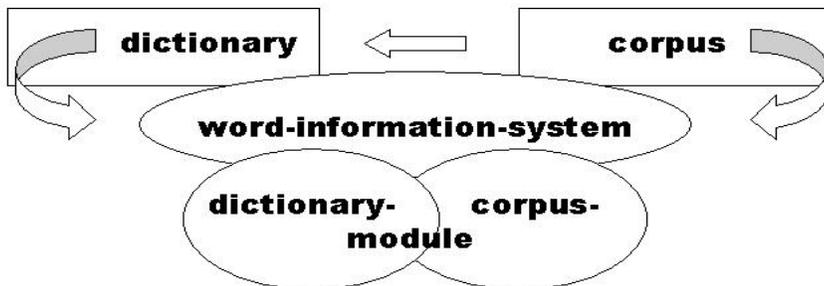[20] Schlaps 2006, 313, gives a prospected period of 135 years to the completion of elexiko.

Figure 1: Hybrid word-information-systems.

A decision whether online dictionaries and corpora should be offered as hybrid systems as suggested here or remain completely separated has still to be made and to be discussed. In any case, from a linguist's point of view, carefully measured convergence is welcome, hybridity enables for usage patterns as shown in (2a) where it was essential to retrieve a word list derived from a dictionary headword list (types) and for patterns as in (2b), where both dictionary component and corpus component (tokens) contribute to answering the problem given in the experiment.

3f Perspectives
In this paper I have attempted to bring together an evaluation of user-pragmatic aspects of two internet-based word-information-systems for German with an evaluation of their applicability for linguistic research. The two examples applied as query-experiments can be considered typical for German, where, for example, in many instances the border between word and phrase is fluent (as, e.g., in example 2b). Processes in word formation are also a prominent feature of German: the first example given (2a) is one which could have been solved with a traditional reverse-dictionary; but such printed databases are not available in any case and they are not as up-to-date and not as complete as data obtained from (good) corpora. Such questions arising from the interest in word-formation, e.g. possible compounds with a base morpheme *-drom* (*Velodrom, Motodrom, Sambadrom*) or others with a similar second element which require a reverese-alphabetical order to be addressed, can best be covered from type-lists derived from (large and diverse) corpora.

The user perspective employed in this paper has necessarily been a limited one, its method being the simulation of a linguist user. Comprehensive user-oriented approaches should attempt to include various other user-groups: an especially interesting species of user for such projects as elexiko and the DWDS could be the L2-learner, which could also be simulated in a manner as presented in this approach. On the other hand, access strategies of non-professional users in my opinion are hardly predictable. Here empirical research is necessary. Simulation can only be the first step towards active (by e.g. questionnaire) and passive (by tracking of user behaviour) user studies which both should be important elements of planning and developing such projects, but hitherto have not received as much attention as they deserve.

The overall aim of this paper was to give impulses for a consideration of user-oriented approaches in the conception of word-information-systems: a possible response by the makers

of the evaluated projects to any observations made above could be that some of the denoted shortcomings are due to the different aims as well as to the unfinished nature of the projects, which in any case is an unconquerable argument; – but, in accordance with the introductory statement of this paper, a project should also focus on the needs and expectations of the users. A succeeding online-project has happy makers as well as happy users.

References

Brown, K. (Ed.) (2006). Encyclopedia *of Language & Linguistics, 2nd edition.* Amsterdam: Elsevier.

Fleischer, W. & Barz, I. (1995). Wortbildung *des Deutschen*. Tübingen: Niemeyer.

Eichinger, L. M. (2000). *Deutsche Wortbildung: Eine Einführung*. Tübingen: Narr.

Geyken, A. & Klein, W. (2001). Projekt "Digitales Wörterbuch der deutschen Sprache des 20. Jh.". *Jahrbuch der BBAW 2000*, Berlin: Akademie-Verlag.

Haß, U. (Ed.) (2005). *Grundfragen der* elektronischen *Lexikographie: elexiko – das Online-Informationssystem zum deutschen Wortschatz.* Berlin/NY: de Gruyter (Schriften des Instituts für deutsche Sprache; 12).

Hunston, S. (2006). Corpus Linguistics. In Brown (2006). Vol. 3, 234-248.

Jucker, A. H. (1994). New Dimensions in Vocabulary Studies: Review article of the Oxford English Dictionary (2nd edition) on CD-ROM. *Literary and Linguistic Computing* 9/2, 149-154.

Klosa, A. & Steffens, D. (2007). Deutscher Wortschatz im Internet: Das Informationssystem elexiko und sein Modulprojekt Neologismen. In H. Kämper & L. M. Eichinger (Eds.): *Sprach-Perspektiven: Germanistische Linguistik und das Institut für Deutsche Sprache* (pp 443-463). Tübingen: Narr (Studien zur Deutschen Sprache: Forschungen des Instituts für deutsche Sprache; 40.

Kramer, U. (2006). German Lexicography. In Brown (2006). Vol. 5, 45-52.

Lemnitzer, L. (2001). Das Internet als Medium für die Wörterbuchbenutzungsforschung . In Lemberg, I., Schröder, B. & Storrer, A. (Eds.): *Chancen und Perspektiven computergestützer Lexikographie. Hypertext, Internet und SGML/XML für die Produktion und Publikation digitaler Wörterbücher* (pp. 247-254). Tübingen: Niemeyer. (Lexicographica S.M. 107). Michaelis, F. (to appear) [conference report]. Tagungsbericht zum Kolloquium "Das *elexiko*-Portal – Präsentation und Diskussion", 10-11 May 2007 at Institut für Deutsche Sprache in Mannheim. *Lexicographica*.

Schlaps, Ch. (2006) [review]. Grundfragen der elektronischen Lexikographie. *elexiko* – das Online-Informationssystem zum deutschen Wortschatz. Hg. v. Ulrike Haß. Berlin/NY: de Gruyter 2005 (Schriften des Instituts für deutsche Sprache 12), 334 Seiten. *Lexicographica* 22/2006, 311-314.

Scherer, C. (2006). *Korpuslinguistik*. Heidelberg: Winter (Kurze Einführungen in die germanistische Linguistik; 2).

Stanulewicz, D. (2007). Polish colour terms referring to blue: A corpus view. In Magnusson, U., Kardela, H. & Głaz, A.: *Further Insights into Semantics and Lexicography* (pp. 87-99). Lublin: Wydawnictwo Uniwersytetu Marii Curii-Skłodowskiej.

Wiegand, H. E. (1987). Zur handlungstheoretischen Grundlegung der Wörterbuch-benutzungsforschung. *Lexicographica* 3, 178-227.

Winhart, H. (2002). *Funktionsverbgefüge im Deutschen: Zur Verbindung von Verben und Nominalisierungen.* Tübingen: Phil. Diss. Retrieved October 20, 2007 from: http://deposit.ddb.de/cgi-bin/dokserv?idn=974495387&dok_var=d1&dok_ext=pdf&filename=974495387.pdf.

[WDG=] *Wörterbuch der deutschen* Gegenwartssprache 6 vols. 1964ff. Berlin: Akademie-Verlag.