Natural Language Processing Techniques for Improved User-friendliness of Electronic Dictionaries

Ulrich Heid
Universität Hildesheim, Germany
heid@uni-hildesheim.de

Abstract

We discuss Natural Language Processing (NLP) tools and techniques which may be used to enhance the user friendliness of electronic dictionaries. Intended properties of electronic dictionaries on which we focus are improved guidance for text production, as well as easy and efficient access to lexical data in text reception dictionaries. In this talk, we focus on those NLP techniques which are mostly available for the major European languages: morphological analysis of inflection and word formation, as well as syntactic analysis. We also address the relevance of a detailed classification and representation of lexical data categories within the dictionary: this is a central prerequisite for any integration of dictionaries and NLP tools. Our discussion is embedded in an interpretation of the claims of the lexicographic Function Theory with respect to user orientation in dictionaries.
Keywords: electronic dictionaries; NLP tools and techniques; user orientation

1 Introduction

In this article, we give a short overview of existing and possible applications of Natural Language Processing (NLP) that could be used to enhance the user friendliness and usability of electronic dictionaries. Enhancing user friendliness in this context means providing better guidance in text production dictionaries, as well as improving access to the data provided by reception dictionaries. In the long term, we envisage integrated lexical information systems that combine a dictionary with a number of Natural Language Processing components.

We first situate our discussion in the framework of our view on the lexicographic Function Theory (cf. e.g. Tarp 2008), and we summarize those aspects of the Function Theory that are directly relevant for the integration of language processing and dictionaries (section 2). As the internal representation of lexical and lexicographical data is a key element in the interaction between the lexicographic and the language processing components, we devote a short section to the issue of data categories and markup (section 3). In section 4, we then discuss existing and likely upcoming language processing devices for text production dictionaries (section 4.1) and for text reception dictionaries (section 4.2). Before we conclude, we address a few general design issues and questions related with the presentation of language processing results to the user (section 5).

2 User Orientation in Dictionaries

Requesting user friendliness of printed and in particular of electronic dictionaries has almost become a common place of lexicographic theory, but also of the advertisements for dictionaries. An example from metalexicography is the lexicographic Function Theory (cf. e.g. Tarp 2008) which places the user and his1 needs in the centre of its reasoning about dictionary design.

2.1 Metalexicographic Viewpoint

As stated by Tarp (2008), dictionaries are to be seen as utility tools. The dictionary is a (possibly network-like) set of structured texts from which a user may be able to extract textual data that allow him, by means of an interactive interpretation, to derive information. The user will go through this interpretation process in response to a need that arises from a non-lexicographic situation. Tarp classifies these needs into several types; the needs immediately relevant to the discussion in this paper are either cognitive or communicative in nature. Cognitive needs arise in situations where the user wants to know about or to learn certain facts, be they about things, concepts or words. Communicative needs arise in (the preparation for) communication activities, i.e. text reception (reading or hearing – and understanding) or text production (writing or speaking – and lexical or grammatical choice). Translation, the revision of texts in a foreign language etc. are also communicative situations, and thus translation towards the mother tongue is a receptive activity, and translation to a foreign language is a production-oriented one.

Dictionaries are supposed to provide appropriate data for users to satisfy needs of the above types. An ideal dictionary, according to the lexicographic Function Theory (henceforth: FT), satisfies exactly one type of need. In terms of FT, the "dictionary function" is to satisfy such a need, and the optimal dictionary is monofunctional. While this ideal is hardly ever commercially viable in printed dictionaries, it can be approximated in electronic dictionaries (cf. e.g. Bergenholtz/Bergenholtz (2011), for an exemplification).

The development of dictionaries, be they printed or electronic, is typically governed by lexicographic processes. These have in the past been exclusively geared towards the production of paper dictionaries, but since the advent of electronic dictionaries (or electronic versions of print dictionaries) they may also be more general (cf. Gouws, to appear), aimed at setting up a repository of lexicographical data that can be used in both an electronic and a print dictionary. We situate the discussion of the use of Natural Language Processing (NLP) tools and techniques in a scenario which is aimed at such a possible double or even multiple use of lexicographic data.

1 For reasons of practicality, we use the masculine form throughout this paper, meant to denote both genders.

2.2 User Orientation in Electronic Dictionaries

In the above sense, we assume that lexicographers collect data to feed more than one dictionary; or, conversely, that not all collected data will show up in one given dictionary. Rather, most dictionary publishers create a broad repository of lexicographic data from which they will select appropriate items for individual dictionaries. This notion is close to the idea of a "mother dictionary" that feeds into several specific dictionaries, a concept introduced into the discussion by R. H. Gouws. Implicitly, the same idea is present in work by the proponents of the Function Theory: to avoid overloading the user, at query time, with (unnecessary) data, they require lexicographers to carefully select the data categories they want to present to the user for a given dictionary function.

While this requirement is very clear at an abstract, metalexicographic level, the way in which it can be satisfied in practice is described in the Function Theory in much less detail (cf., however, the tables given by Tarp 2008: 75-77).

In a scenario where electronic dictionaries are to be produced from an electronic data collection, providing users with data appropriate for a given need involves filtering the contents of the data collection. In the spirit the well-known distinction between lexicographic data description and lexicographic data presentation, filtering has the following aspects:

(1) selection of data categories relevant for a given dictionary function;
(2) selection of presentational properties appropriate for the targeted user public, in terms of the ordering of microstructural items, of their layout, metalanguage, presentational devices, and of the provision of appropriate access routes to the data.

Both of the above selections depend crucially on the following factors:

(1) the dictionary function;
(2) the pre-existing knowledge of users, in terms of the language (or languages) dealt with in the dictionary, as well as of general aspects of dictionary use (Tarp 2008) or of the use of online information tools;
(3) possibly the complexity of the language phenomena described in the targeted dictionary article(s).

The illustration in figure 1 schematically summarizes the relationship between data repository, filtering and user-oriented dictionary versions.

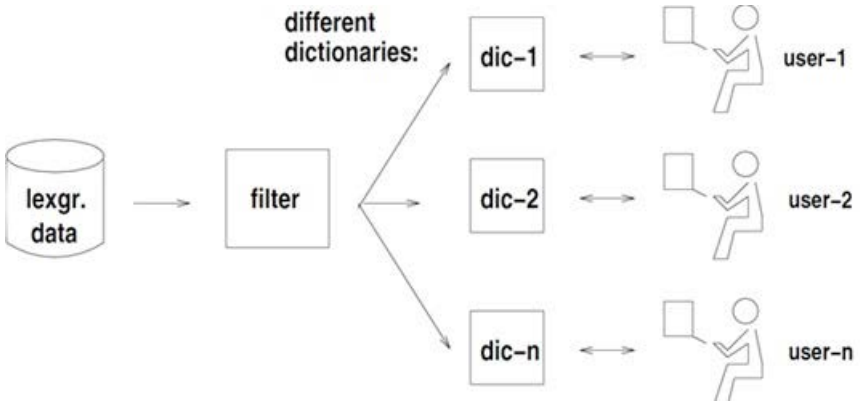

Figure 1: Scenario of the production of user oriented dictionaries: data repository – filtering – monofunctional dictionaries.

The definition of the above mentioned filtering criteria, as well as of the selected presentational devices is in principle part of the dictionary concept ("Wörterbuchplan", in Wiegand's and Gouws's terms); it is similar in nature to specifications of a piece of software; the selection of data categories is a contents-related specification, while the definition of presentational devices and of access routes is mainly a matter of the rendering of individual data categories for printed or on-screen presentation. The specification of access to the data is mainly conditioned by the lexicographer's assumptions about the user's pre-existing knowledge of the domain or language treated in the dictionary. For example, it is plausible to assume that users in need of collocational data, for text production, will know the base of the collocation (in the sense of Hausmann 2004), and will search for an appropriate collocate to express a given idea. Thus a sort of onomasiological access would be preferred: one that allows the user to search for expressions of ideas around a base concept, irrespective of whether these ideas are expressed by collocations, compounds or single words (cf. Giacomini 2012)

This difference in access is illustrated, in figure 2, below, for a printed dictionary (or a print-like electronic one) with sample data from the OCDSE (Oxford Collocations Dictionary for Students of English) for the lemma advance: for text production, the data are sorted according to OCDSE's principles (right side of fig. 2), i.e. per reading of the base, with subdivisions per syntactic model and semantic groups (cf. also Heid/Zimmermann 2012); for text reception (left side), a semi-integrated microstructure is suggested, with a section on the readings of the base, followed by an alphabetical listing of collocational adjectives (and, later in the article, but not shown in figure 2, of collocational verbs and nouns); in an electronic dictionary, reception-oriented access would be more flexible, i.e. from both elements of the collocation, as well as from the collocation as a whole.

Reception

- Readings
  (1) [military] forward movement
  (2) development
  (3) amount of money
- Typical adjectives
  - Allied etc. (cf. German etc) (1)
  - big (= considerable) (2)
  - cash (3)
  - considerable (= big) (2)
  - dramatic (2)
  - German (cf. Allied, etc) (1)
  - great (2)
  - important (1)
  - large (3)
  - notable (2)

Production

Reading 1: forward movement [military]

- ADJ + advance
  - [speed] rapid ~
  - [agent] German ~, Allied ~, etc.
- V + advance
  - [make] make an ~ on X
  The regiment made an advance on the enemy lines

Reading 2: development (often in the plural)

- ADJ + advance
  - [amount] considerable ~, big ~, substantial ~, dramatic ~, enormous ~, great ~, spectacular ~, tremendous
- V + Advance
  - [make] make ~ es (in/on) [plural]

Reading 3: amount of money

- ADJ + advance
  - [quantity] small ~, large ~ - [type] cash ~
- V + Advance
  - [provide] give so, an ~, pay so, an ~
  The university pays me an advance for thir business trip.

Figure 2: Microstructures for easy access to collocations: for text reception (left) vs. text production (right).

3 Data Representation and User Orientation

For an electronic dictionary scenario which involves a central data collection, monofunctional (or at least function-related) dictionaries and the appropriate filtering techniques, a cornerstone of successful implementation is a detailed classification of the available lexicographic data. If, for example, no distinction is made between collocations and idiomatic expressions, and if bases and collocates of collocations (in the above mentioned sense of Hausmann 2004) are not distinguished and marked up, it will be hard if not impossible to provide appropriately differentiated access to collocations vs. idioms. If, for example, both are classified as "multiword expressions", it will be hard to decide which items will go into a text production dictionary and which ones into a text reception dictionary. Most lexicographers would however agree that collocations are relevant for text production (but not needed – with the exception perhaps of what Grossmann/Tutin 2003 would call "opaque collocations", such as FR peur bleue – for text reception), while idiomatic expressions would need to be semantically explained in a text reception dictionary, but would rather not be described, for instance, in a learner-oriented text production dictionary.

If the classification of data categories is central, so is the functional markup of different data categories in the central repository used by a publisher. Some authors call this repository a "database" (e.g. Bergenholtz 2011). This may be technically adequate in the implementation described by Bergenholtz

(2011), but in the general case, this repository need not be a database in the technical sense; publishing houses may also use XML-based data models, or a representation within a content management system, or any other implementation: what counts is that different data categories are distinguished and identifiable. There are examples of publishing houses which use the fine-grained data classification of their data repository to "extract" dictionaries for certain functions and user-groups, without much need for adding new data. Such detailed data categorization and "markup" is also necessary if certain subsets of the lexicographical data available to a publisher are to be provided for the purposes of Natural Language Processing (NLP).

A note of caution may be in order here, with respect to the abovementioned example of collocations and idioms. Some lexicographers might rightly assume that users will not be able to distinguish between the two types of multiword expressions, and claim that they don't need to (cf. e.g. Tarp 2008). This is certainly an appropriate viewpoint, but it does not distinguish the internal representation if lexic(ographic)al data from the presentation of such data to the user. If the abovementioned assumptions about the differences between idioms and collocations, in terms of data selection and access, are correct, an optimal presentation of such data to the user will clearly have to rely on the distinction between the two types of multiword expressions, and on an appropriate markup of each multiword item contained in the data repository. In other words: an optimal presentation of dictionary contents to the user is (trivially) dependent on an adequate internal classification and representation if this contents.

3 Natural Language Processing Tools in Support of User Orientation

In our discussion, so far, Natural Language Processing (= NLP) tools have not been mentioned; a sensible level of user-friendliness can, as has been shown above, be reached without NLP technology, by adhering to good practice in data category classification and markup. A legitimate question is thus what the added value of computational linguistic technology is, in terms of a surplus of user-friendliness of the dictionary. Here, we understand user orientation in a wide sense; it is meant, here, to include aspects of enhanced usability of electronic dictionaries, such as improved access to lexicographic data, individualized support according to the user's pre-existing knowledge, or the availability of information (on demand) which goes beyond the amount of material encoded in the data repository underlying the dictionary, e.g. by means of the presentation of corpus data.

We will address this question in the following by first analyzing monolingual dictionaries for text production, then dictionaries for text reception. We will not discuss the use of NLP techniques for the provision of corpus data to the lexicographer, i.e. corpus analysis and query tools, such as, for instance, those embodied in the "Sketch Engine" (Kilgarriff et al. 2004). Such tools are essentially aimed at the

lexicographer, and we will show in which way the end user may profit from other kinds of corpus analysis tools.

4.1 NLP Tools for Text Production Dictionaries

For text production, especially in a foreign language, a rich microstructure is necessary, for example one which explains to the user for each treatment unit which morphological forms it has, which syntactic patterns it follows, or which collocations it enters into. All these properties involve lists of options from which a user may want to select in a text production situation; while syntax and collocations are idiosyncratic (Hausmann 2004 talks of "coded combinatorics") and need thus be listed individually, morphological forms are often more regular and may be provided to the user by means of a morphological generator or of a list of inflection forms. The latter is limited and may require regular updates by the lexicographer, whereas a morphological generator may provide the advantage of comparatively easy ways of extending its coverage. In any case, offering users of a production dictionary on-demand access to inflection forms is certainly very useful.

4.1.1 Access to Corpus Data

Another possible use of NLP techniques in text production dictionaries has to do with the much discussed facilities that give the user access to corpus data, from a dictionary entry, (cf. e.g. Asmussen 2013; Heid et al. 2012; Tarp 2012). It has been claimed that links from the dictionary to corpus data or to the internet provide users with data about language in use which can serve as a model for the users' own text production. Such data thus serves the need for checking one's own formulation hypotheses against putatively "standard" usage.

Asmussen (2013) discusses the issues related with the realization of such links; using the German DWDS dictionary as an example, he shows that the mere coexistence of dictionary and corpus data in one portal is not sufficient to provide adequate service to users. Asmussen has examples of lemmas present only in one of the two sources of information, and he discusses the impossibility of linking corpus data, given today's technology, to readings of a dictionary entry, at least in a large-scale high-quality way. Asmussen's best example of the linking of dictionary and corpus data is from the domain of collocations. In fact, the portal of ordnet.dk provides direct access to usage examples for collocations, with example sentences retrieved from Korpus 2000, the current Danish corpus underlying the DDO dictionary (Den Danske Ordbog, cf. ordnet.dk). To activate the link (in the sense of an information-on-demand offer), the user has to press a button next to the collocation item in the dictionary. Technically, this activates a query in the corpus which is created from the (text form of) the collocation item (cf. Heid et al. 2012).

This facility could be further enhanced if the corpus were preprocessed at a more advanced level of linguistic analysis (e.g. by means of (flat) syntactic analysis), and if the query were made more sensitive to potentially ambiguous corpus sentences. For example a search that starts from the collocation

give + resultater (EN "produce/lead to results") provides many relevant examples, e.g. gav betydelige resultater ("gave important results"), gav de ønskede resultater ("gave the intended results"); but it also provides gav 10.3 km/l til resultat ("gave 10.3 km/l as a result"), which is not an example of the searched collocation.

Other, similar facilities involve the retrieval of contexts for specialized terms in specialized texts, e.g. in the intranet of a company, or lists of cooccurrents of items (for the user to choose from) sorted by association strength, as they are provided within Verlinde's ILT tool (URL: https://idp.kuleuven.be/idp/view/login.htm).

On the basis of syntactically analyzed and annotated texts, the same device could also be offered for syntactic subcategorization. Parsed corpus data tend to identify subjects, objects, prepositional complements, verb-dependent clauses or infinitivals in each analyzed sentence; this is true for dependency parsing, which has reached, at least for several European languages, a degree of maturity makes its use in the intended context possible (cf. e.g. Bohnet 2010). Syntactic valency is, as mentioned above, an important property of lexical predicates (verbs, adjectives and nationalizations) that must be learned by foreign language learners. Illustrating valency in full sentences has the advantage of showing the user not only an abstract indication, but also concrete instanciations of it. This principle has been followed, very successfully, in the ELDIT dictionary (Elektronisches Lernerwörterbuch Deutsch-Italienisch, http://eldit.eurac.edu/), where the authors have provided the user with four complementary types of indications for the syntactic construction of verbs (cf. Abel 2002):

(I) a formula of the type "someone suggests something to someone";

(II) example sentences for each pattern;

(III) on-demand indications of the involved grammatical functions (e.g. "object" for "something" in the above example);

(IV) on-demand highlighting of the respective phrases in the example sentences, when the user points the mouse to an element of the formula (i): if, for example, the user points the mouse to "to someone" in the above example, not only the grammatical function (indirect object) is displayed, but also the respective stretch of the example sentence is highlighted.

In ELDIT, these devices are applied to prefabricated examples, i.e. a closed list of verbs and example sentences for these. By use of NLP tools, such a device could be made dynamic, i.e. provided on demand by the user, on the basis of a pre-analyzed (dependency-parsed) corpus and extraction tools for syntactic patterns. If the corpus is big enough and adequately annotated, also frequency data for individual valency constructions and lists of the most prominent fillers of valency relations could be provided. Finally, as Engelberg et al. (2012) have shown, the different possible valency constructions of verbs are used differently in different genres or text types: not all possible patterns are equally frequent in all kinds of texts; if notions like "genre" or "text type" are applied to the annotation of a corpus which is exploited to offer examples for valency patterns explained in a dictionary, such valency preferences by genre or text type can be made visible.

All this may appear futuristic to some readers; it is, however, only dependent on two conditions: an adequately detailed inventory of valency patterns in the dictionary, and good quality corpus parsing. The device would allow users to get real-text models of syntactic constructions, from which they could take inspiration for their own text production.

Such devices are in principle thinkable for all those linguistic properties of lexical items that can unambiguously be identified in a (syntactically parsed) corpus. With adjective+noun-collocations and, to a lesser extent, verb+object- and verb+subject-collocations, this is well possible; the same holds for syntactic valency, for the contextual use of terms from a specialized language; but it does not yet so for lexical semantic properties. Some partial results could be obtained if additional resources are used, e.g. WordNets that would support a search for word combinations and the pertaining sentences according to semantically defined sets of lexical items. Again, what is needed as a prerequisite, are appropriate classifications of the lexicographic data, mappable onto the classifications annotated in the corpus. In all cases, the combination of lexical items and targeted linguistic properties acts as search criteria for corpus data extraction.

Instead of lexical items and their linguistic properties, also pairs of translational equivalents from an electronic bilingual dictionary may be used as search constraints, in this case on parallel corpora; Verlinde's ILT tools provide access to the Europarl corpus (URL: http://www.statmt.org/europarl/), but they use only the source language item as a search criterion, in the hope of providing the user in this way a broad range of equivalence candidates; for very advanced users, and especially for those who are used to work with parallel corpora, this may provide indeed new insight. A more modest, though perhaps more focused (and thus easier to use) version of such search would be one that retrieves example sentence pairs for a given equivalent pair from the dictionary; it may then on demand also provide sentence pairs that do not contain the equivalents mentioned in the dictionary entry, as a complement of information.

4.1.2 Lexico-grammatical Guidance

With the above mentioned provision of example sentences for a given lexical property of an item from the dictionary, one problem mentioned by Asmussen (2013) still remains: for the dictionary user, the relationship between the text he is in the process of producing, and the sentences retrieved from the corpus, may still be rather indirect; the examples may illustrate the behaviour of the searched item, but they will still not necessarily provide an exact solution to the actual text production problem which the user is confronted with, as the other lexical items he intends to use are absent from the retrieved examples.

While the transfer between the corpus examples and the upcoming text of the user may be relatively simple for most European languages, it is much harder in languages with massive rule-based morphosyntactic variation, where lexical choice and grammatical choice interact in more complex ways. Examples of such situations are provided by the South African Sotho and Nguni languages. The morphosyntactic complexity of the noun class system of these languages, of their concordial and prono-

minal morphemes, as well as of their tense and mood systems interferes with issues of lexical choice that depend on semantic selection criteria. Examples are discussed, among others in this conference, by Bosch/Faaß (2014) and Prinsloo et al. (2014); they analyze possessive constructions in Zulu (type: the medicine of this doctor) and subjects, objects and relative clauses in Northern Sotho (type: the boy who helped the woman), respectively, from the viewpoint of English → Zulu or English → Sotho learner's dictionaries.

Both model the respective phenomena in an NLP tool that implements the morphosyntactic (agreement) rules of the language and interacts with a dictionary whose nominal entries are classified by noun classes and whose verbal entries can be inserted into the grammar of the constructions under analysis. Bosch/Faaß's (2014) system can operate in two modes: one that provides a translation from English to Zulu of the intended possessive construction, and one that in addition explains to the user which construction and agreement rules have been used. A next step could be a system that allows the user to produce his own solution and which then proposes modifications where necessary. Prinsloo's system is not yet implemented but intended to provide similar optional guidance: if, for example, the user plans to construct a Northern Sotho subject-verb-object sentence where the object is not expressed by a noun (phrase), but by a concord, the following situations may occur:

(I) the user may know the appropriate concord: the system will check the appropriateness and confirm it;

(II) the user may know the noun which he wants to express by the concord, but not the concord itself: the system will retrieve the noun class of the item (from its standard dictionary), identify the appropriate concord (from morphosyntactic tables) and propose the appropriate concord, possibly with additional information about the underlying grammatical facts;

(III) the user may only know the English equivalent of the nominal he intends to construct as an object: the system then retrieves the appropriate Sotho noun from a bilingual dictionary and then proceeds as explained under (ii).

In both cases, the objective is text production guidance for learners of the foreign language; the proposed solutions combine some amount of NLP with a well-structured dictionary. Such combined systems may be counted among online language learning systems, or among e-dictionaries. Irrespective of how they are classified, they combine lexicon and (partial) grammar data.

4.2 NLP Tools for Text Reception Dictionaries

While the function of NLP tools in the context of text production goes from a source of inspiration (through the selection of appropriate example sentences from corpora) to guidance in issues of lexico-grammatical choice, NLP tools function mostly as access tools in text reception dictionaries. Text reception starts from an existing text and aims at detecting its meaning and possibly other properties. Access support tools ease the user's retrieval of the right dictionary entry and the right indications within this entry. This kind of support starts with inflectional morphology and may involve

word formation, syntax and possibly, at least to some extent, multi-word items and semantics. In all cases, the basic function of the tools is to analyse a word(form), possibly in the context of the sentence the user is reading, and to relate it to an entry or ideally even a reading in the dictionary.

For inflectional morphology, such devices work relatively well and are quite established: if the user enters a word form, the dictionary relates it to the appropriate base form and displays the entry of the pertaining lemma. Such interfaces exist in several online dictionaries, and they may either depend on large lists of inflected forms (related with the appropriate lemmas), or on morphological analysers.

However, similar devices may be used also for word formation: when Bergenholtz/Johnsen (2005) analyzed the log files of their Danish Internet Dictionary (Dansk Netordbog), they discovered that a considerable number of items searched by users, but not found in the dictionary, were word formation products. In Germanic languages, compounding is so productive that a standard monovolume dictionary cannot cover even a small portion of the items found with a non-trivial number of occurrences in a corpus. The same holds, at least to some extent, for derivation products: these also will likely not be covered in full in standard monovolume dictionaries. If the dictionary is meant for text reception, it may even reasonably adopt a policy of focussing on semantically non-transparent compounds and derivations, i.e. on those whose meaning needs to be explained beyond a simple recall of the morphological structure, e.g. because they have idiosyncratic meanings. In such a case, no space may be left for the treatment of transparent compounds.

A morphological word formation analyzer may be useful in such a situation, as it would be able to split a compound that is not part of the nomenclature of the dictionary into its components and guide the user to their entries in the dictionary. For derivation, even generic paraphrases may be given, as is the case in the DériF system for French (URL: http://www.cnrtl.fr/outils/DeriF/). Alternatively, a structural or morpheme decomposition hypothesis may be given, as in the canoo tools (URL: http://www.canoo.net/). In all cases, the user would get partial information in reply to his query, thus at least some guidance towards the analysis of the word formation product. Often, a recall of the morphological structure and links to the dictionary entries of the components may help users understand complex word.

While the above functions are in general seen independent from the context (see below for a discussion), any kind of tool intended for guidance of text readers towards an appropriate entry in the dictionary will inevitably be confronted with ambiguity and context-based disambiguation. This starts with categorial homographs (EN: can: modal verb or noun, cf. Bothma/Prinsloo 2013: 176) or forms which are homographous with items from other word classes (thought: participle or noun, ibid. 174) and goes all the way up to polysemy and reading distinctions. While the latter problems are in the general case still not solvable, categorial homographs of different types can be disambiguated fairly well on the basis of word class tagging and/or syntactic analysis. For the major European languages, tools for these functions are available, also as web services that would allow for an on-the-fly treatment, as suggested by Bothma/Prinsloo (2013: 187 ff.).

If syntactic (dependency) analysis is available, a variant of the search for examples discussed above, in section 3.1, could be applied; if a syntactic analysis of a sentence can be produced, at least the verb and its potential complements can be extracted from the analysis result and matched against the list of valency patterns offered by a dictionary. In many cases, this would reduce the number of readings which the user would need to consult in order to find the meaning of the verb used in the sentence he is reading. Obviously, not always all arguments of a predicate are explicitly mentioned in a sentence, which might increase the number of syntactic readings that have to be looked up by the user in such a situation.

Another type of contextual analysis has been available since the 1990s, already. It deals with (idiomatic) multiword expressions; the objective is to provide the user with the (part of a) dictionary entry that explains the multiword expression, when the user clicks on any of the words that make up the expression. Due to the high level of lexical specificity this device works quite well for idioms (cf. Seretan/Wehli 2013). In combination with a syntactic analysis of the text which the user is working on, this facility could be extended to collocations as well.

5 NLP Tools as a Part of Lexical Information Systems

In the two preceeding sections, we have listed a few simple devices based on NLP technology that could enhance the user friendliness of electronic dictionaries. A number of issues should however be discussed in this context which concern more general aspects of user interaction.

An important first aspect concerns the problem of the quality of the tool output; it can not be guaranteed that the linguistic analysis underlying the integrated tool is correct in all cases. The more NLP components are involved, the more possibilities are there for errors to occur in the processing chain. Thus a setup where, for example, the text being read by the user is syntactically analyzed and then searched, in order to find the appropriate dictionary entry for a given item, may provide fully correct results only in a certain percentage of cases (we would assume at least in three quarters of the cases).

The dictionary developer and the users should be aware of this situation, and such a device should be offered as experimental (cf. Tarp 2012 on this issue in the context of corpus data provision). It would need to be presented to the dictionary user as being fully automatic and not cross-checked by the lexicographer. A warning in the user interface would be appropriate in such a case (e.g. 'N.B.: the following results were automatically produced and have not been checked by lexicographers'). This type of warnings is regularly produced by the canoo morphology system when analyses are displayed which have not been cross-checked by canoo's lexicographers.

Furthermore, the results should be shown in a part of the interface that is clearly recognizable as a part of the dictionary or of the lexical information system. Early realizations, especially of tools that link dictionary and corpus, did not make clear enough to the user that the corpus data shown on

screen were meant as an additional service of the dictionary (cf. Bank 2012 on an early version of the Base lexicale du français).

Another, related issue concerns the status of NLP-based services within a dictionary system. In our view, they should always be information-on-demand services: the user should have the possibility to explicitly decide in favour (or against) the use of the NLP-based service. In dictionaries with (personalized) user profiles, this choice may be an element of the user profile.

Finally the design of the overall integrated system should also be governed by the principles of user-oriented dictionary design: for each NLP component to be integrated, the lexicographer should assess which user need (and thus: dictionary function) it satisfies.

6 Conclusion

We have shown a few devices for the enhancement of text production and text reception dictionaries that are based on Natural Language Processing tools. While morphology systems or morphological tables are almost a standard component of current electronic dictionaries, this is much less so with tools for syntactic analysis, and technologies for semantic processing are still in an experimental phase, although some are very promising (cf. e.g. Cook 2014).

We have tried to show that a detailed classification of the data categories contained in an electronic dictionary (or in the data repository underlying it) is a major requirement for any work with NLP tools; this is due to the fact that these data categories (and a common "understanding" about them, between NLP tools and lexical repository) are the interface between the two components.

In our view, syntactic dependency - parsing has reached a degree - of maturity, at least for some European languages, which should allow for its experimental - and perhaps also productive - integration into lexical information systems of the kind discussed here. Prototypes of such systems should be built and tested with users.

If lexicographers join forces with NLP experts, and if they jointly produce integrated systems that present an added value, chances are good that users will accept these systems. Since people are particularly demanding with respect to the quality of language tools (and since dictionaries have a high reputation in this respect), offering integrated services as information-on-demand seems to be an adequate solution.

7 References

Abel, A. (2002). Darstellung der Verbvalenz in einem elektronischen Lernerwörterbuch Deutsch – Italienisch (ELDIT). Neue Medien, neue Ansätze. In: Braasch, A. et al. (eds.): EURALEX-2002 Proceedings, pp. 413-418.

Asmussen, J. (2013). Combined products: Dictionary and corpus. In: Gouws, G.H., Heid, U., Schweickard, W., Wiegand, H.E. (eds.) (2013). Dictionaries. An International Encyclopedia of Lexicography. Volume 5.4. pp. 1081-1090.

Bank, C. (2012). Die Usability von Online-Wörterbüchern und elektronischen Sprachportalen. Information - Wissenschaft & Praxis. Volume 63/ 6. pp. 345–360. Accessed at: http://www.degruyter.com/view/j/iwp.2012.63.issue-6/iwp-2012-0069/iwp-2012-0069.xml?format=INT [28/05/2014].

Bergenholtz, H. (2011). Access to and Presentation of Needs-Adapted Data in Monofunctional Internet Dictionaries. In: Fuertes-Olivera, P. A., Bergenholtz, H. (eds.): e-Lexicography: The Internet, Digital Initiatives and Lexicography, pp. 30-53.

Bergenholtz, H., Bergenholtz, I. (2011). A Dictionary Is a Tools, a Good Dictionary Is a Monofunctional Tool. In: Fuertes-Olivera, P. A., Bergenholtz, H. (eds.): e-Lexicography: The Internet, Digital Initiatives and Lexicography, pp. 187-207.

Bergenholtz, H., Johnson, M. (2005). Log Files as a Tool for Improving Internet Dictionaries. In: Hermes, (34), pp. 117-141.

Bohnet, B. (2010). Very High Accuracy and Fast Dependency Parsing is not a Contradiction. In: Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), Beijing, China.

Bothma, T. J. D., Prinsloo, D. J. (2013). Automated dictionary consultation for text reception: a critical evaluation of lexicographic guidance in Kindle e-dictionaries. Lexicographica: International Annual for Lexicography, Volume 29, pp. 165-198.

Bosch, S., Faaß, G. (2014). Towards an integrated e-dictionary application - the case of an English to Zulu dictionary of possessives. 16th EURALEX International congress, 15-19 July, 2014, Bolzano/Bozen (EURAC).

Canoo. Accessed at: http://www.canoo.net/ [27/05/2014].

Cook, P., Rundell, M., Han Lau, J., Baldwin, T. (2014). Applying a Word-sense Induction System to the Automatic Extraction of Diverse Dictionary Examples. 16th EURALEX International congress, 15-19 July, 2014, Bolzano/Bozen (EURAC).

DériF. Accessed at: http://www.cnrtl.fr/outils/DeriF/ [27/05/2014].

ELDIT – Elektronisches Wörterbuch Deutsch – Italienisch. Accessed at: http://eldit.eurac.edu/ [27/05/2014].

Engelberg, S., Koplenig, A., Proost, K., Winkler, E. (2012). Argument structure and text genre: cross-corpus evaluation of the distributional characteristics of argument structure realizations. Lexicographica: International Annual for Lexicography, Volume 28, pp. 13-48.

Fuertes-Olivera, P. A., Bergenholtz, H., Nielsen, S., Niño Amo, M. (2012). Classification in lexicography. The concept of collocation in the Accounting-Dictionaries. Lexicographica: International Annual for Lexicography, Volume 28, pp. 293-307.

Fuertes-Olivera, P. A., Bergenholtz, H. (eds.) (2011). e-Lexicography: The Internet, Digital Initiatives and Lexicography. London: Bloomsbury.

Giacomini, L. (2012). An onomasiological dictionary of collocations: mediostructural properties and search procedures. Lexicographica: International Annual for Lexicography, Volume 27, pp. 241-267.

Gossmann, F., Tutin, A. (eds.) (2003). Les collocations: analyse et traitement. Amsterdam : De Werelt.

Gouws, R. H. (2013). Aspekte des lexikographischen Prozesses in Print- und Onlinewörterbüchern. To appear. OPAL (IdS Mannheim).

Hausmann, F. J. (2004). Was sind eigentlich Kollokationen? In: Steyer, K. (ed.): Wortverbindungen – mehr oder weniger fest. Jahrbuch des Instituts für Deutsche Sprache 2003. Berlin/New York: De Gruyter, pp. 309-334.

Heid, U., Prinsloo, D. J., Bothma, T. J. D. (2012). Dictionary and corpus data in a common portal: state of the art and requirements for the future. Lexicographica: International Annual for Lexicography, Volume 28, pp. 269-289.

Heid, U., Zimmermann, J. T. (2012). Usability testing as a tool for e-dictionary design: collocations as a case in point. EURALEX-2012 Proceedings, Oslo, Norway, pp. 661-671.

ILT Tools. Accessed at: https://idp.kuleuven.be/idp/view/login.htm [27/05/2014].

Kilgarriff, P. R., Smrz, P., Tugwell, D. (2004). The Sketch Engine EURALEX-2004 Proceedings. Lorient, France, July: pp. 105-116.

Prinsloo, D. J., Bothma, T. J. D., Heid, U. (2014). User support in e-dictionaries for complex grammatical structures in the Bantu languages. 16th EURALEX International congress, 15-19 July, 2014, Bolzano/Bozen (EURAC).

Seretan, V., Wehrli, E. (eds.) (2013). Context-sensitive look-up in electronic dictionaries. In: Gouws, G.H., Heid, U., Schweickard, W., Wiegand, H.E. (2013). Dictionaries. An International Encyclopedia of Lexicography. Volume 5.4. pp. 1046-1053.

Trap-Jensen, L. (2013). Researching lexicographical practice. In: Jackson, H. (ed.): Bloomsbury Companion to Lexicography, pp. 35-47. London: Bloomsbury Academic.

Tarp, S. (2008). Lexicography in the Borderland between Knowledge and Non-knowledge: General Lexicographical Theory with Particular Focus on Learner's Lexicography. Tübingen: Max Niemeyer.

Tarp, S. (2012). Online dictionaries: today and tomorrow. Lexicographica: International Annual for Lexicography, Volume 28, pp. 253-267.