··· *Papers*

> **Reuse of Lexicographic Data for a Multipurpose Pronunciation Database and Phonetic Transcription Generator for Regional Variants of Portuguese**

ASHBY, SIMONE AND FERREIRA, JOSÉ PEDRO

1 – *Computational Lexicography and Lexicology*

Among the benefits of a flexible and modular lexical database are: the facility of building new modules from existing ones, the reuse of lexicographic data to both enhance the user experience and achieve NLP aims, the time saved in accomplishing these objectives, and the economy that comes from minimizing redundancy (van der Eijk, Bloksma, and van der Kraan, 1992). LUPo, or the Portuguese Unisyn Lexicon, is one of the first *speech*-dedicated applications to take full advantage of a collection of lexical resources as the basis for a text-to-speech system. Consisting of a pronunciation lexicon and rule system for generating accent-specific phonetic transcriptions for Portuguese, LUPo circumvents the cost of producing high-quality phonetic transcriptions by hand, while attracting a wider pan Lusophone audience to the lexical database in which it resides, and providing the research community with a vast resource of Portuguese accent data for evaluating speech applications and testing theories.

> **From the Definitions of the *Trésor de la Langue Française* to a Semantic Database of the French Language**

BARQUE, LUCIE; NASR, ALEXIS AND POLGUÈRE, ALAIN

1 – *Computational Lexicography and Lexicology*

The *Definiens* project aims at building a database of French lexical semantics that is formal and structured enough to allow for a fine-grained semantic access to the French lexicon—for such tasks as automatic extraction and computation. To achieve this in a relatively short time, we process the definitions of the *Trésor de la Langue Française informatisé* (TLFi), enriching them with an XML tagging that makes explicit their internal organization (roughly, *genus* and *differentiae*) and enhancing the components with semantic labels that explicit their role in the definition. There is, to our knowledge, no existing broad coverage database for the French lexicon that offers to researchers and NLP developers a structured decomposition of the meaning of lexical units. Definiens is an ongoing research that will hopefully fill this gap in the near future.

> **Morphosyntactic Lexica in the OAL Framework: Towards a Formalism to Handle Spelling Variants, Compounds and Multi-word Units**

BLANCAFORT, HELENA; COUTO, JAVIER AND SENG, SOMARA

1 – Computational Lexicography and Lexicology

The creation and maintenance of lexicographic resources are labour-intensive tasks. In this paper we present SylLex, a formalism to encode morphosyntactic lexica and how it is used in the OAL framework, a tool to aid the linguist to create and maintain such resources in an industrial context. The aim is to have an intuitive and easy-to-use formalism, SylLex, implemented in a user-friendly and ready-to-use tool so that a linguist without previous experience is able to work effectively after a short training of one or two hours.
The paper is organised as follows. First, we present the SylLex formalism. SylLex organises the lexical and inflectional information by dividing the lexicon in three components: lemma, inflection paradigms and patterns. Thus, the linguist can better manipulate the data, assure consistency and correctness, and maintain the lexicon by modifying inflection paradigms instead of dealing with scripts to directly modify lexicon entries in text files. In addition to this, the system takes advantage of the notion of inheritance between different inflection paradigms of a same word category. Furthermore, we present how different kinds of variants of forms like geographic and spelling variants are encoded in SylLex. Moreover, we discuss the problem of defining and delimiting compound and multi-word units, and explain how they are stored in the lexicon. Finally, we draw conclusions on the advantages and disadvantages of the formalism and OAL and mention further work.

> **TTC: Terminology Extraction, Translation Tools and Comparable Corpora**

BLANCAFORT, HELENA; DAILLE, BÉATRICE; GORNOSTAY, TATIANA; HEID, ULRICH; MECHOULAM, CLAUDE AND SHAROFF, SERGE

1 – Computational Lexicography and Lexicology

The need for linguistic resources in any natural language application is undeniable. Lexicons and terminologies play indeed a central role in any machine translation tool, regardless of the theoretical foundations upon which the machine translation tool is based (e.g. statistical machine translation or rule-based machine translation). The EU project TTC ('Terminology Extraction, Translation Tools and Comparable Corpora') aims at leveraging machine translation tools, computer-assisted translation tools, and terminology management tools by automatically generating bilingual terminologies from comparable corpora in several European Union languages (English, French, German, Latvian

and Spanish), as well as in Chinese and Russian. The TTC project will integrate developed and existing tools in an online platform including a tool to compile and handle comparable corpora, as well as a terminology management tool. The platform will be based on Web Services and will use reputable open solutions such as UIMA (Unstructured Information Management Architecture) and EuroTermBank.

> The Past Meets the Present in Swedish FrameNet++

BORIN, LARS; DANÉLLS, DANA; FORSBERG, MARKUS; KOKKINAKIS, DIMITRIOS AND TOPOROWSKA GRONOSTAJ, MARIA

1 – Computational Lexicography and Lexicology

The paper is about a recently initiated pilot project which aims at the development of a Swedish framenet as an integral part of a larger lexical resource, hence the name 'Swedish FrameNet++' (SweFN++). The SweFN++ project has four main goals: (1) to 'revitalize' a number of existing lexical resources and integrate them into a multi-faceted lexical resource for language technology (LT) applications, in the process enriching the individual resources using semi-automatic methods; (2) to construct a Swedish framenet (SweFN) and make it part of the integrated resource; (3) to develop a methodology and workflow which makes maximal use of LT and other tools in order to minimize the human effort needed to build the resource; and (4) to release the resource under an open content license. The above goals are also of great significance for lexicological research and computational lexicography, as a SweFN will lend relevant support in bringing to light semantic relations implicit in word meanings. The theoretical assumptions elaborated by the Berkeley FrameNet make up the backbone of the SweFN resource, which will pay special attention to compounds and multi-words expressions when used as target lexical units or frame elements. In this article, we present an inventory of free electronic resources with a focus on their role in the semi-automatic acquisition and population of Swedish frames. After a brief overview of Swedish resources, we reflect on attempts to recycling and linking lexical data in a semi-automatic manner and report on our work in progress, which can be followed at <http://spraakbanken.gu.se/swefn/eng/>.

> Encoding Attitude and Connotation in wordnets

BRAASCH, ANNA AND PEDERSEN, BOLETTE S.

1 – Computational Lexicography and Lexicology

The Danish wordnet, DanNet, though part of the global WordNet family, contains some information types that are not generally provided in

wordnets such as qualia roles and *connotation* of words. Connotation is seen as the set of associations implied by a lexeme in addition to its primary, literary meaning; it is evoked by one (or more) particular feature of the entity referred to and suggests attitudes, emotions and opinions like admiration or disapproval. Lexemes with a connotation have an observable pragmatic effect in texts making them *subjective* or *opinionated*.

In the paper, we discuss the relevance of connotation information in lexicons for computational applications in general and present the set of encoded semantic information exemplified by empirical data. We focus on a particular ontological type of entities, namely *human*s with the focus on selected hyponyms of *person* that are encoded with a connotation value and discuss the prototypical properties evoking *positive* or *negative* connotations. The qualia structure based approach enables to encode both the prevalent, connotation evoking features and prototypical activities of the person.

The material encoded with connotation so far consist of 650 nouns and comprises a male, a female and a gender-neutral group, thus it lends itself to comparative examinations concerning the distribution of connotation evoking features and polarity distribution within each individual group and between the groups as well. One of the most striking observations says that (in our material) the negative connotation polarity is predominant; the most important feature of female persons seems to be their positive appearance, and a general disparaging attitude dominates as regards the conduct and manners of male persons.

> The DANTE Database (Database of ANalysed Texts of English)

CONVERY, CATHAL; Ó MIANÁIN, PÁDRAIG; Ó RAGHALLAIGH, MUIRIS; ATKINS, SUE; KILGARRIFF, ADAM AND RUNDELL, MICHAEL

1 – Computational Lexicography and Lexicology

This database (www.webDante.com) was designed and created for Foras na Gaeilge by the Lexicography MasterClass and their 15-strong team led by Valerie Grundy (Managing Editor); textflow is managed by Diana Rawlinson (Project Administrator). The corpus of 1.7 bn words of current English, custom-built in 2007, was queried using the Sketch Engine (www.sketchengine.co.uk/), and the database was compiled in IDM's Dictionary Production System (DPS: www.idm.fr). The present volume contains a fuller description of this project (Atkins, Kilgarriff and Rundell *Database of ANalysed Texts of English (DANTE): the NEID database project*) and of its use in a bilingual dictionary (Convery, Ó Mianáin and Ó Raghallaigh *Covering all bases: Regional Marking of material in the New English-Irish Dictionary).*

The 95,000 or so DANTE entries cover approximately 50,000 headwords and 45,000 compounds, idioms and phrasal verbs, using over 40 datatypes. The lexical entry is subdivided into lexical units, each a sense of a single- or multi-word lemma. Almost every linguistic fact recorded is accompanied by full corpus sentences illustrating its use in text. Apart from the definitions and the corpus-derived example sentences, all the significant information is machine-retrievable. Functionality demonstrated here includes simple and complex searches over various combinations of datatypes and the automatic insertion of empty translation fields for use in dictionary building. .

DANTE was created as the initial stage of compilation of the *New English-Irish Dictionary*. Its long-term potential is much more far-reaching: it offers publishers world-wide a comprehensive launchpad for bilingual dictionaries with English as the source language or the draft stage of a learners' dictionary of English; a source of updating material for an existing dictionary, etc. It offers software developers, universities and other research institutions a resource for improved word sense disambiguation, the creation or enhancement of online lexicons, and other uses in software applications such as machine-assisted translation, information retrieval systems, etc. More details from info@webDante.com.

> From a Bilingual Transdisciplinary Scientific Lexicon to Bilingual Transdisciplinary Scientific Collocations

DROUIN, PATRICK

1 – Computational Lexicography and Lexicology

Most linguistic studies dealing with the lexicon of scientific corpora are interested in subject area lexicon or terminology which leads to a general lack of description of the other types of lexical items contained in these corpora. The main exception to the previous statement is the work being done in the area of specialized language teaching like the studies of Coxhead (1998, 2000). In most cases, as pointed out by Tutin (2007), the lexicon itself is not what is being studied.

We consider that the lexicon used in scientific writings can be divided into three categories. The first one is the common basic lexicon, which includes function words such as determiners, auxiliary verbs and conjunctions, and content words of the general language. The second category is the transdisciplinary lexicon that includes abstract verbs such as *to think* or *to consider* and abstract nouns such as *idea*, *factor* and *relation*. It also includes a methodological lexicon that refers to the abstract lexicon used for the description of scientific activities and scientific reasoning. Examples of lexical items one would find in this category are *hypothesis*,

data and *approach*. The last category of lexicon found in scientific writings is subject specific terminology, which refers to all concepts used in a particular domain.
Our study here is focused on the second category, which we called Transdisciplinary Scientific Lexicon (TSL), and its behavior in scientific writings. The main goal of this paper is to test the idea that we can start from a raw bilingual scientific corpus and automatically build a list of bilingual transdisciplinary scientific collocations around the lexical items from the second category described above (TSL).

> iLEX, a general system for traditional dictionaries on paper and adaptive electronic lexical resources

ERLANDSEN, JENS

1 – Computational Lexicography and Lexicology

Dictionaries are different: the purpose, the language(s) covered, authors co-operation during production, and the traditions and styles have asked for many solutions and editorial rule sets. The needs of individual projects can be covered in two ways: The specific system is customized at the programming level (maybe with reuse of existing modules); the general system is customized at a higher level that does not require programming skills.
iLEX is a general system integrating spell checking, structural and lexical help, lists, workflow, smartEditing, advanced graphical statistics, separated metadata, change tracking, fast powerful searching, alphabetization, tilde functions, element sorting and uniqueness control, and more for Windows, Linux and Mac. For more details, visit www.emp.dk. Based on full Unicode, XML schemas, Namespace, ISO Schematron, Xquery, Xpath and XSL 2.0 iLEX is a secure choice for the future.
Being 100 percent XML conforming, iLEX may be used for a broader range of texts: not only TEI and related standards as MENOTA, but any of the emerging application standards, e.g. DITA.
Single sourcing an indispensable aspect of XML, is a strong feature of iLEX: XSL publication on any platform is supported as well as ready-to-publish ilEX Comunico applications for internet and mobile phones using Symbian60 and Android.
Lately, online publishing has changed towards more adaptive approaches for two aspects. Adaptation of user interface layout and functionality to ensure higher user satisfaction is a must when financed by ads. Adaptation of content based on composition of both existing lexical resources and resources dynamically generated from computational analysis of language use on the internet is a way to keep costs within control.

This asks for new tools and methods. Lexical Information Mapping Architecture (LIMA) which will be sketched out as a part of this presentation may form the ground for a new standard for adaptive lexical work. It will relate to other standards as XML, TEI, DITA, SCORM, and ISO 1951 and more.
For more information about iLEX and LIMA visit www.emp.dk

> **Corpus Exploitation Strategies for the Lexicographic Definition Task**

FELIU, JUDIT; GIL, ÀNGEL; PEDEMONTE, BERTA AND GUIRADO, CRISTINA

1 – Computational Lexicography and Lexicology

The main goal of this paper is to formalize and to present some guidelines helping the lexicographic definition task. The research is applied to the corpus query procedures in order to retrieve some refined results that benefit the definition writing up process as much as possible. The paper will briefly introduce the three main language resources (LR) involved in the authors' daily linguistic job, that is, the Catalan descriptive dictionary built on the basis of a Catalan corpus, the corpus itself and the Catalan main dictionaries repository normally looked up. The focus of the paper will be put on the improvement of the strategies followed so far in order to use the corpus query system in an efficiently oriented manner as far as the descriptor selection and the extrinsic part of the definition fulfilment is concerned. The use of set of patterns will allow retrieving certain kind of information for each type of unit defined. Data retained will be more precise and the results are proved to be useful for maintaining coherence among different team members and also, and probably most important, among different but semantically related types of words defined in the descriptive dictionary.

> **Mit einem Klick zu vielen Möglichkeiten: *www.deutsches-rechtswoerterbuch.de***

FRIELING, STEFANIE

1 – Computational Lexicography and Lexicology

The *Deutsches Rechtswörterbuch (DRW)* is a historical dictionary (based at the *Heidelberger Akademie der Wissenschaften*) which documents and describes the vocabulary of the historical legal language of 7th to 19th century German. *German* here refers to the West Germanic language family which includes amongst others Old Frisian, Old Anglo Saxon, Old Saxon, Langobardic, Low German und High German. Besides the words directly referring to juridical expressions (such as *Gericht* or *Prozess*) the dictionary also includes many words of the common language if they

occur in a legal context (e.g. in edicts or contracts). This conceptual design makes the DRW an important tool for a diverse range of historic research, e.g. cultural and social history, history of law, history of economy or linguistics.
Until now, 11 of the 16 planned volumes have been published. The majority of the nearly 100.000 dictionary articles is freely available online, too. The web version of the DRW, however, is much more than just the digital pendant of the printed reference book: Via the website www.deutsches-rechtswoerterbuch.de the dictionary user has access to a wide range of options to use the dictionary as a large information system.
Based on the software FAUST – also used by the lexicographers for their daily work – four databases (the dictionary itself, the sources, the digitised sources, the full text archive) can be searched online.
During the software presentation the use of the DRW web version as a powerful research tool will be demonstrated. By means of a variety of possible enquiries it will be shown how the user benefits from the interconnectedness of the FAUST databases and the integration of external information resources (e.g. other important historical dictionaries or digitised primary sources).

[Note: the author did not submit a full text version of this software demonstration]

> The *Louvain EAP Dictionary (LEAD)*

GRANGER, SYLVIANE AND PAQUOT, MAGALI

1 – Computational Lexicography and Lexicology

In our software demonstration, we describe a web-based English for Academic Purposes dictionary-cum-writing aid tool, the *Louvain EAP Dictionary (LEAD)*. The dictionary is based on the analysis of c. 900 academic words and phrases in a large corpus of academic texts and EFL learner corpora representing a wide range of L1 populations. The dictionary contains a rich description of non-technical academic words, with particular focus on their phraseology (collocations and recurrent phrases). Its main originality is its customisability: the content is automatically adapted to users' needs in terms of discipline and mother tongue background. Another key feature of the *LEAD* is that is makes full use of the capabilities afforded by the electronic medium in terms of multiplicity of access modes (Tarp 2009). The dictionary can be used as both a semasiological dictionary (from lexeme to meaning) and an onomasiological dictionary (from meaning/concept to lexeme) via a list of typical rhetorical or organisational functions in academic discourse (cf. Pecman 2008). It is also a semi-bilingual dictionary (cf. Laufer &

Levitzky-Aviad 2006) as users who have selected a particular mother tongue background can search lexical entries via their translations into that language.
The *LEAD* is designed as an integrated tool where the actual dictionary part is linked up to other language resources and learning tools. It is a hybrid dictionary (cf. Hartman, 2005) that includes both a dictionary-cum-corpus and a dictionary-cum-CALL component. As regards direct corpus access, the *LEAD* innovates by giving access to discipline-specific corpora rather than generic corpora.
While the current version of the tool is restricted to some disciplines and mother tongue backgrounds, its flexible architecture allows for further customisation (other L1 background populations, other disciplines, other languages).

> **One Structure for Both Monolingual and Bilingual Dictionaries Converting a Large Number of Different Dictionaries to a Single XML Format**

GROOT, HANS DE AND MASEREEUW, PIETER

1 – *Computational Lexicography and Lexicology*

Van Dale converted the source databases of all its dictionaries to a type of XML mainly designed to capture the function of its elements, rather than their formatting. We found that we could apply the same principles consistently to various types of dictionaries (monolingual, bilingual) and capture all content within a single XML structure. The new structure reduces the time needed for our production processes and for database maintenance. This article reports on our findings during the conversion and the principles we applied.

> **Corpus-derived data on German multiword expressions for lexicography**

HEID, ULRICH AND WELLER, MARION

1 – *Computational Lexicography and Lexicology*

We show a parsing-based architecture for the extraction of German verbal multiword expressions. It uses dependency parsing as a preprocessing step, allows us to extract syntactic patterns of arbitrary form from the parsed data, and comprises a relational database where each extracted multiword occurrence is stored along with the sentence it is extracted from, and with a number of morphosyntactic and syntactic features. These features serve (i) for an automatic decision about the likely idiomatization of the candidate under review, and (ii) in later lexicographic work to get

a clear picture of lexicographically relevant linguistic properties of the selected candidates.
We use dependency-parsed text, because this allows us to find non-adjacent multiwords and to use subcategorization knowledge to identify e.g. verb + object pairs more reliably than on the basis of ourface patterns.
The extraction results illustrate the potential of the tools; we can identify morphosyntactic preferences in collocations (these often indicate idiomatization), longer collocational or idiomatic structures (where e.g. the core elements and possible modifies can be clearly distinguished), lexical variation in idioms, as well as certain specific features of collocations or idioms (e.g. preferences for negation).
As all data are stored in a database, which supports a variety of generalization steps, it is in principle possible to prepare different layouts (i.e. presentations and selections) of dictionary entries, for different user groups and user needs.

> **Dictionary Building based on Parallel Corpora and Word Alignment**

HÉJA, ENIKŐ

1 – Computational Lexicography and Lexicology

The paper describes an approach based on word alignment on parallel corpora, which aims at facilitating the lexicographic work during dictionary building. This corpus-driven technique, in particular the exploitation of parallel corpora, proved to be helpful in the creation of bilingual dictionaries for several reasons. Most importantly, a parallel corpus of appropriate size guarantees that the most relevant translations are included in the dictionary. Moreover, based on their translational probabilities it is possible to rank translation candidates, which ensures that the most likely translation candidates are ranked higher. A further advantage is that all the relevant example sentences from the parallel corpora are easily accessible, thus alleviating the selection of the most appropriate translations from possible translation candidates. Due to these properties the method is particularly apt to enable the production of active or encoding dictionaries

> **Using a Dictionary Production System to impose a WordNet on a Dictionary. A Software Presentation**

HVELPLUND, HOLGER AND ØRSNES, ALLAN

1 – Computational Lexicography and Lexicology

In this demonstration, we will make a practical presentation of how a semantic web – a WordNet – can be imposed on a dictionary using a Dictionary Production System. Using a WordNet structure makes a lot of

sense as there are already several WordNets for different languages which can be used freely.
We will demonstrate how we first impose the structure in a rather crude way allowing for lots of ambiguities and then use the various tools in the Dictionary Production System to target the dubious areas and refine them through manual intervention, either as part of a usual editing of a new edition or in a separate run.
Finally we will demonstrate how the data can be extracted and made available to a Dictionary Publishing System.

> Working with the web as a source for dictionaries of informal vocabulary

JANSSON, HÅKAN

1 – Computational Lexicography and Lexicology

Informal vocabulary, e.g. slang, jargon and other forms of expression that are particular to different types of small or closed groups, is usually suppressed in writing that has passed an editorial process. This means that most of the corpora used for lexicographical are lacking in this area. Today, however, the Internet has given us new possibilities to tap into to the flow of colloquial and informal language. The aim of this presentation is foremost to give a brief account of how the Internet could be 'harvested' for the purpose of creating corpora which include substantial amounts of informal language, and secondly, how to use these (in this case Swedish and Icelandic) corpora to gather candidates for headwords with informal markings such as ***coll.***, ***slang***, and the like. The topic of evaluation of results of this kind of work will also be touched upon. The work here presented has been done utilizing Sketch Engine, and strategies employed in using that tool are thence also accounted for.

> Building Russian Word Sketches as Models of Phrases

KHOKHLOVA, MARIA

1 – Computational Lexicography and Lexicology

Without any doubt corpora are vital tools for linguistic studies and solution for applied tasks. Although corpora opportunities are very useful, there is a need of another kind of software for further improvement of linguistic research as it is impossible to process huge amount of linguistic data manually. The Sketch Engine representing itself a corpus tool which takes as input a corpus of any language and corresponding grammar patterns. The paper describes the writing of Sketch grammar for the Russian language as a part of the Sketch Engine system. The system gives information about a

word's collocability on concrete dependency models, and generates lists of the most frequent phrases for a given word based on appropriate models. The paper deals with two different approaches to writing rules for the grammar, based on morphological information, and also with applying word sketches to the Russian language. The data evidences that such results may find an extensive use in various fields of linguistics, such as dictionary compiling, language learning and teaching, translation (including machine translation), phraseology, information retrieval etc.

> **A Quantitative Evaluation of Word Sketches**

KILGARRIFF, ADAM; KOVÁŘ, VOJTĚCH; KREK , SIMON; SRDANOVIC, IRENA AND TIBERIUS, CAROLE

1 – Computational Lexicography and Lexicology

A word sketch is an automatic corpus-derived summary of a word's grammatical and collocational behaviour. Word sketches were first prepared in 1999 for the compilation of the Macmillan English Dictionary for Advanced Learners (Rundell 2002). They have since been integrated into the Sketch Engine corpus query tool (Kilgarriff et al 2004), prepared for fifteen languages, and used on a large scale for lexicography by a number of publishers. We are frequently told how impressive they are and how little they miss – but we would like a more rigorous assessment.
We describe a formal, quantitative evaluation of word sketches, from a user perspective, for four languages (Dutch, English, Slovene, Japanese), with the critical question being 'is the collocation suitable for inclusion in a published collocation dictionary'. For each language, we inspected twenty collocates for each of forty-two headwords. In each case two thirds or more of the collocations were of publishable quality.

> **Semantic Relations in Cognitive eLexicography**

KREMER, GERHARD AND ABEL, ANDREA

1 – Computational Lexicography and Lexicology

Whereas dictionary design has traditionally been guided by the results of dictionary use research, recent approaches in lexicographic research are strictly user-centred. We support the idea of integrating empirical cognitive evidence into this type of research, thus fruitfully exploiting it for both, the selection (and subsequently presentation) of lexical data and the acquisition of such data from corpora. Focusing on the extraction of semantic relations to be illustrated in electronic learners' dictionaries, we analyse the results of two behavioural experiments on the production as well as the perception of semantic relations. The main

goal of the experiments was to determine which relations are cognitively salient in speakers' minds. With the objective of developing a method to automatically extract cognitively salient semantic relations from corpora, we describe and discuss findings of the first analyses conducted on composite part relations. In future this might serve as a basis for the elaboration of new strategies aimed at enriching lexical databases and dictionaries.

> The IDM Free Online Platform for Dictionary Publishers

LANNOY, VINCENT

1 – Computational Lexicography and Lexicology

Printed dictionaries have built a genuine identity over the years. Lexicographers work for renowned publishers according to specific rules and processes; distribution channels are well-organized and efficient at delivering to educational or public markets. The emergence of new actors, exclusively focused on the Web, is a major upheaval as they deliver large corpora to a worldwide audience. Those Pure Players are now dominating the online dictionary market not only in terms of audience but also by establishing their own brands, independent of existing print brands.

These new actors bring their own vision of what an online dictionary should be. This presents a great opportunity for the industry to rethink the way dictionaries are written and published, inspired by the distinctive strengths of the Internet as a medium which call for clarity of the information, easiness of the service, and above all, intrinsic value of linguistic, i.e. lexicographic data.

Our experience, built through day-to-day management of several major free online dictionary websites, demonstrates the strong draw of dictionary content. Since dictionary websites encompass a very broad spectrum of the language and make it available for free on the Internet, users discover online dictionaries by very diverse means. Their distinct paths to a dictionary reflect their different interests in the content, and also their different expectations for the content delivered.

Making dictionary data amenable to favourable placement in search engines, for searches made in many languages, requires close involvement of lexicographers. These lexicographers must adapt to a process of creating entries for dynamic display on screen in addition to static display in print; understanding the impact of Search Engine Optimization (SEO) on entry structure; integrating a rich network of hyperlinks and making use of non-textual media to enrich their lexical content. Lexicographers are in the spotlight of the digital paradigm!

Quality of the content and publishers' care over data play a key role in

building user loyalty and depth of visit on the Website. On average, in a language learning context, we observe that visits last between 5 and 7 pages, providing the publisher with the opportunity to be in contact with its users for several pages. The question is *to do what?* For the moment, most of the dictionary websites are dead ends: a user enters for one or several definitions and leaves though his needs or interests can be much deeper. He may require course books, vocabulary lists, exercises for learners, novels, reference content, etc. Affiliation models help propose not only the publisher's own content but complementary contents, products or services coming from partners. We are currently successfully experiencing with a partner the efficiency of an up-sales model based on dictionary free entries. Dictionary content is not only an efficient attraction point but plays also the role of a *user qualification filter for targeted up-sales*. Dictionary is an intermediary between a query and a targeted product.

> **Constructing a Constructional MWE Lexicon for Psycho-Conceptual Annotation: An Evaluation of CPA and DUELME for Lexicographic Description**

LUDER, MARC AND CLEMATIDE, SIMON

1 – Computational Lexicography and Lexicology

The German JAKOB lexicon provides a basis for the coding of patient narratives and is currently extended in the direction of a phraseological and construction-grammar resource. For this purpose, we will compare two formalisms for the representation of multiword expressions (MWE): The Dutch Electronic Lexicon of Multiword Expressions (DuELME, Grégoire 2009) and the verb patterns from Corpus Pattern Analysis (CPA, Hanks 2008). We are looking for a representation format which is human-readable, and equally adapted for natural language processing (NLP). The JAKOB lexicon is implemented in the OLIF format and currently contains 7000 entries. The MWEs investigated are verbal phraseologisms and originate from the corpora of three different clients, consisting of a total of more than 400 transcribed sessions.

The narrative analysis method JAKOB is a tool for investigating everyday stories from psychotherapy transcripts (Boothe, 2004). Stories are annotated on the basis of our predefined psycho-conceptual coding system represented in the lexicon. JAKOB allows formulating hypotheses about the client's conflicts, the analysis of the discourse being one component thereof.

DuELME is an NLP lexicon project which encodes MWE descriptions in a theory- and implementation-independent way. Every MWE is an instance of a construction class with elements including morpho-

syntactic parameters. CPA patterns represent semantic properties for the elements of a (verbal) construction, whereas syntactic properties are represented in the JAKOB lexicon by the subcategorization frames (Satzmuster) of Wahrig (2007). We are implementing an additional lexicon property 'bauplan' which is formally constructed as a combination of the DuELME component list, the Wahrig subcategorization frame and semantic information out of the CPA-pattern. Because this structure is difficult to read for the lexicographer, it is generated automatically and can be hidden from the user, but is available for NLP tasks.

> **Une nouvelle ressource lexicographique en ligne: le Petit Larousse Illustré de 1905**

MANUÉLIAN, HÉLÈNE

1 – Computational Lexicography and Lexicology

Cet article présente une nouvelle ressource lexicographique ancienne mise à disposition sur Internet : le Petit Larousse Illustré de 1905. Faisant suite à des œuvres de plus grande ampleur et de plus grande renommée (le dictionnaire critique de Féraud, le dictionnaire de Nicot, les différentes éditions de celui de l'Académie, etc.), le Petit Larousse Illustré de 1905, bien plus modeste que ses prédécesseurs – en volume tout au moins – a été numérisé et sera mis en ligne prochainement.
L'intérêt de la mise en ligne d'une telle ressource réside dans sa nature. Il s'agit d'un petit dictionnaire illustré, et la présence d'images est importante. Par ailleurs, il est le premier d'une série de dictionnaires grand public, ce qui le rend fondamental dans l'histoire de la lexicographie.
L'informatisation s'est déroulée en plusieurs phases, de façon à permettre une interrogation fine du dictionnaire. Les différents éléments des articles du dictionnaire ont été décrits et listés, puis balisés en XML selon les standards décrits dans la proposition 5 de la TEI. Le texte a ensuite été balisé automatiquement grâce à des programmes écrits en langage Python contenant des expressions régulières. Le balisage s'est déroulé en trois passes, chacune exploitant le résultat de la précédente.
Le résultat de l'informatisation est une base de données lexicales riche qui permet à l'utilisateur deux sortes de consultations : il peut choisir de faire une interrogation plein texte. Dans ce cas, le résultat apparaîtra avec les images associées aux articles répondant à sa requête. L'utilisateur peut aussi faire une recherche avancée, c'est-à-dire n'interroger qu'un seul champ de l'article du dictionnaire (vedette, prononciation, information grammaticale, étymologie, définitions, définitions encyclopédiques, renvois, proverbes, exemples, expressions figées). Seules les requêtes sur la vedette permettent l'affichage des images.

> **Getting Synonym Candidates from Raw Data in the English Lexical Substitution Task**

McCARTHY, DIANA; KELLER, BILL AND NAVIGLI, ROBERTO

1 – Computational Lexicography and Lexicology

Distributional similarity provides a technique for obtaining semantically related words from corpus data using automated methods that compare the contexts in which the words appear. Such methods can be useful for producing thesauruses, with application to work in lexicography and computational linguistics. However, the most similar words produced using these methods are not always near synonyms, but may be words in other semantic relationships: antonyms, hyponyms or even looser 'topical' relations. This means that manual post-processing of such automatically produced resources to filter out unwanted words may be necessary before they can be used. This paper evaluates the performance of distributional methods for finding synonyms on the English Lexical Substitution Task, a lexical paraphrasing task where it is necessary to generate candidate synonyms for a target word and then select a suitable substitute on the basis of contextual information. We examine the performance of distributional methods for the first step of generating candidate synonyms and leave the second step of choosing a candidate on the basis of context for future work. A number of automated distributional methods are compared to techniques that make use of manually produced thesauruses. We demonstrate that while the performance of such automatic thesaurus acquisition methods is often below manually produced resources, precision can be greatly increased by using two automatic methods in combination. This approach gives precision results that surpass methods that exploit manually constructed resources for the same task, albeit at the expense of coverage. We conclude that such an approach to increase the precision of automatic methods to find near synonyms could improve the use of distributional methods in lexicography.

> **What WordNet does not know about selectional preferences**

MĚCHURA, MICHAL BOLESLAV

1 – Computational Lexicography and Lexicology

Selectional preferences are the tendencies of words to co-occur with other words that belong to certain semantic types. In this paper, I will investigate how closely these corpus-attested preferences correspond to WordNet. For example, for all possible direct objects of *cancel*, is there a single category (or a union of several categories) in WordNet that subsumes them, and only them? Selectional preferences manifest themselves in authentic

texts and can be revealed through corpus analysis. I will introduce an experimental tool I have built which attempts to do this automatically by aligning corpus-extracted lists of collocates (for example a list of the direct objects of *cancel*) with WordNet. The strength of this method is that it can discover and name selectional preferences automatically, but its weakness is that it can only do so when WordNet contains a suitable category. We will see that WordNet often lacks a category (or even a union of several categories) that fully corresponds to an attested selectional preference – for example, there is no category in WordNet that includes all the kinds of events that can be direct objects of *cancel* (*meeting, wedding, concert* etc.) but excludes those that cannot (*accident, sunset, invention* etc.).

> OWID – A dictionary net for corpus-based lexicography of contemporary German

MÜLLER-SPITZER, CAROLIN

1 – Computational Lexicography and Lexicology

The *Online-Wortschatz-Informationssystem Deutsch* (OWID; Online German Lexical Information System) is a lexicographic Internet portal for various electronic dictionary resources that are being compiled at the Institute for the German Language (Institut für Deutsche Sprache, IDS). The main emphasis of OWID is on academic lexicographic resources of contemporary German. Presently, the following dictionaries are included in OWID: a dictionary of contemporary German called *elexiko*, a dictionary of neologisms, a small dictionary of collocations, and a discourse dictionary covering the lexemes that establish the discourse about 'guilt' in the early post-war era 1945-1955. In the near future (2010/2011), several additional dictionaries will be published in OWID: a Textbook of German Communication Verbs, a Valency Dictionary of German Verbs, two further discourse dictionaries – one about the 'democracy' discourse around 1968, the other covering the keywords of the German reunification 1989/1990. Moreover, 300 entries from a corpus-based project on proverbs will be integrated into OWID. Thereby, OWID is a constantly growing resource for academic lexicographic work of the German language.

Altogether, OWID is a special kind of dictionary portal owing to its content and its design, namely the integration of the various dictionaries, the access possibilities and the presentation features. With OWID, we try to establish a dictionary net where the different resources are jointly accessible not only by headwords, but also on the microstructural level. Prerequisite for these common access- and navigation-possibilities across the various dictionaries is the same concept for the lexicographic data model which we put into practice in OWID. Data from all dictionaries

in OWID are structured according to a tailor-made, fine-granular, XML-based data model. In this data model, similar content is modelled similarly, dictionary related differences are preserved.
The main tasks for the future are to enhance OWID with further dictionary resources, to improve the inner access structures so that they exhaust the possibilities of the data model, and to customize the layout of the dictionaries as well as the search options according to the user's needs.

> OBELEX – the 'Online Bibliography of Electronic Lexicography'

MÜLLER-SPITZER, CAROLIN AND MÖHRS, CHRISTINE

1 – Computational Lexicography and Lexicology

Digital or electronic lexicography has gained in importance in the last few years. This can be seen in the growing list of publications focusing on this field. In the OBELEX bibliography (http://www.owid.de/obelex/engl), the research contributions in this field are consolidated and are searchable by different criteria. The idea for OBELEX originated in the context of the dictionary portal OWID, which incorporates several dictionaries from the Institute for German Language (www.owid.de). OBELEX has been available online free of charge since December 2008.
OBELEX includes articles, monographs, anthologies and reviews published since 2000 which relate to electronic lexicography, as well as some relevant older works. Our particular focus is on works about online lexicography. Systematically evaluated sources are relevant journals like *International Journal of Lexicography*, *Lexicographica*, *Dictionaries*, *Lexikos*; furthermore *Euralex-Proceedings*, proceedings of the *International Symposium on Lexicography* in Copenhagen as well as relevant monographs and anthologies. Information on dictionaries is currently not included in OBELEX; the main focus is on metalexicography. However, we are working on a database with information on online dictionaries as a supplement to OBELEX.
All entries of OBELEX are stored in a database. Thus, all parts of the bibliographic entry (such as person, title, publication or year) are searchable. Furthermore, all publications are associated with our keyword list; therefore, a thematic search is possible. The subject language is also noted.
With this type of content, the OBELEX bibliography supplements in a useful way other bibliographic projects such as the printed 'Internationale Bibliographie zur germanistischen Lexikographie und Wörterbuchforschung' by H. E. Wiegand (Wiegand 2006/2007), the 'Bibliography of Lexicography' by R. R. K. Hartmann (Hartmann 2007), and the 'International Bibliography of Lexicography' of Euralex (cf. also DeCesaris and Bernal 2006). OBELEX differs from all these bibliographic

projects by its strong focus on electronic lexicography and its ability to retrieve bibliographic information.

> **Towards Semi-Automatic Dictionary Making**
> **Creating the Frequency Dictionary of Hungarian Verb Phrase Constructions**

PAJZS, JÚLIA AND SASS, BÁLINT

1 – Computational Lexicography and Lexicology

The paper describes the lexicographical aspects of creating a frequency dictionary by a semi-automatic process. The bulk of the work is made by task specific software. The output of the program is then manually checked, corrected and filtered. The result is a collection of the most frequent Hungarian verb phrase constructions (VPCs), illustrated by corpus examples. This is a corpus driven dictionary, based on the 187,6 million word synchronic Hungarian National Corpus (http://corpus.nytud.hu/mnsz) which was analyzed by a series of programs. Its output is a set of XML format draft entries, which were then hand validated and edited by lexicographers. The dictionary contains the most frequent Hungarian verbs along with their most typical syntactic constructions. At the current phase of the project we decided to collect the most frequent constructions only: their absolute frequency had to be more than 250. The dictionary contains roughly 2300 entries and 6500 VPCs. Each construction is illustrated by a corpus example. The verbal entries are presented in alphabetical order primarily. Different kinds of indices are also included in the printed version. The users of this dictionary envisaged to be mainly linguists, working on Hungarian grammars, lexicographers working on bilingual dictionaries and last but not least: advanced level learners of Hungarian, who want to expand their knowledge on the Hungarian nominal verbal collocation relationships. The dictionary is planned to be published both in printed and electronic format.
Parts of the algorithm used for this project could be applied to produce other dictionaries, all the more so, as some of them are actually language independent. It is also highly cost effective: both the programming and the lexicographic work required one person year each.

> **Developing GiGaNT, a lexical infrastructure covering 16 centuries**

RUITENBERG†, TILLY; DOES, JESSE DE AND DEPUYDT, KATRIEN

1 – Computational Lexicography and Lexicology

GiGaNT is a new INL initiative which sets out to develop a computational lexicon (lexical database) covering 16 centuries of Dutch language. This

means that all lexical data of the dictionaries, corpora and computational lexica of the Institute for Dutch Lexicology (INL) will be stored into a central database, functioning both as computational lexicon and central infrastructure for the maintenance of lexical data. Dictionaries, corpora and this computational lexicon are all part of the Dutch Language Bank (DLB).

The immediate incentive to develop GiGaNT was the need for a diachronic computational lexicon, to serve both as a link between texts and dictionaries in the DLB and as a solid infrastructure for other, similar lexical data at the INL. The GiGaNT lexicon will be used for text or corpus annotation, facilitating the retrieval and investigation of the annotated texts.

Integration of existing material into GiGaNT and its subsequent adaptation to enable it to function within computational applications will be a huge step towards another aim: the systematic screening of the complete Dutch word stock for 'gaps' in lexicographic description. This applies to both neologisms and hitherto undescribed historical words.

Users will benefit from the possibility to link from word forms in running text to lexicographical definitions in the INL dictionaries. Researchers, who now only have access to separate collections, will benefit as well: in the future they will have one single starting point for their searches and one single basis from which to develop new lexical material. GiGaNT will also give expert users better access to the lexical data maintained by the INL. The infrastructure will function as a database which will be accessible to API's and as a 'service' that enables researchers to compare their data with GiGaNT and eventually to contribute their own material to GiGaNT.

> **Electronic Dictionary and Dictionary Writing System: how this duo works for dictionary user's needs (ABBYY Lingvo and ABBYY Lingvo Content case)**

RYLOVA, ANNA

1 – Computational Lexicography and Lexicology

The main idea we present in this paper is that using special markup in dictionary writing system and having appropriate functionality in electronic dictionary software we can achieve new results in satisfying most important needs of dictionary user. We describe the core functionality of the ABBYY Lingvo Content dictionary writing system and some features of ABBYY Lingvo electronic dictionary software that presents the dictionaries made in DWS. Then we show how the dictionary

data can be used in text translation scenario and how DWS and electronic dictionary work together to meet the user's needs in translation and text analyzing.

Besides the core functionality ABBYY Lingvo Content DWS includes

– embedded in DWS interface user-friendly entry filtration system. Lexicographer doesn't need to know any special query language – just tick boxes in filtration window tabs;
– also embedded in interface tool for dictionary comparison and merge;
– visual markup of changes – you can always compare any two versions of dictionary entry and see what was added, deleted, changed or restored to their earlier versions.;
– possibility of working with many dictionaries (2 and more) in one window, editing their entries simultaneously.

ABBYY Lingvo *electronic dictionary* has been developed since 1989 and nowadays it is used by 7 million users worldwide.

One of the basic dictionary user's need is to find the appropriate translation for the word he met in a text (text reception) or translate a word from their mother tongue to a foreign language. Here we will describe how the electronic dictionary software works to satisfy the text reception need. One of the most challenging task for a dictionary producer – to help the dictionary reader find a good translation and all the relevant information about the word. This task can be done well if a lexicographer puts relevant markup for a dictionary entry in DWS and electronic dictionary has a proper functionality to process this markup and a good interface to show the result of this processing to dictionary user.

> **The Cornetto database: Semantic issues in linking lexical units and synsets**

VLIET, HENNIE VAN DER; MAKS, ISA; VOSSEN, PIEK AND SEGERS, ROXANE

1 – *Computational Lexicography and Lexicology*

Cornetto is a lexical semantic database that combines the Dutch Wordnet (Vossen (1998)) and the Referentie Bestand Nederlands (Van der Vliet (2007)). The Dutch Wordnet (DWN) is similar to the Princeton Wordnet for English (Fellbaum (1998)), and the Referentie Bestand Nederlands (RBN) includes frame-like information as in FrameNet (Fillmore, Baker, Sato (2004)) as well as information on the combinatorical behaviour of word meanings. The combination of the lexical resources has resulted in a rich relational database that may improve natural language processing technologies.

An important aspect of combining the resources is the alignment of the lexical units (LU's) and the synsets. Automatic alignment of RBN and DWN resulted in an initial version of the Cornetto database. This version has

been further extended both automatically and manually. The resulting data structure is stored in a database that keeps separate collections for LU's (mainly derived from RBN), synsets (derived from DWN) and, in addition, a formal ontology (SUMO/MILO, see Niles and Pease (2001)). These 3 semantic resources represent different viewpoints and layers of linguistic and conceptual information. The resulting resource is freely available for research in the form of an XML database.
In this contribution, we will concentrate on the semantic information in Cornetto. We will discuss the differences in the perspective on semantics in the LU's and synsets and we will give a brief overview of the differences with regard to semantic information. The merging of the two resources resulted in very rich semantic database. However, combining lexica with different perspectives on semantics causes specific problems in the alignment of LU's and synsets and leads to findings that shed light on the organization of meaning in the lexicon.

> **¿Lo que necesitan es lo que encuentran? Reflexiones a propósito de la representación de los verbos en los diccionarios de aprendizaje del español**

BERNAL, ELISENDA AND RENAU, IRENE

2 – The Dictionary-Making Process

The verb is one of the most analysed parts of speech in lexicographical research and, as a result, the tendency of establishing and putting into practise microstructure models that include more and more grammar appears to be consolidated. With respect to Spanish foreign learners (ELE) dictionaries, this tendency is still in an initial stage.
We believe it is necessary to pay more attention to users, in order to provide them with this grammatical information, which is required for production. In this sense, we present the results of an experiment with intermediate-advanced students of ELE, to determine if the *Diccionario de español como lengua extranjera* (DAELE, *Spanish Dictionary for Foreign Learners*) that we are developing, in fact satisfies users' needs in this respect. DAELE is an online dictionary that is fully based on corpus analysis is aimed at advanced learners.
We tested two groups of ELE from the Universitat Pompeu Fabra (Barcelona, Spain). In both cases, a control group was used. The test consisted of two exercises, a composition task and a questionnaire in which participants were asked to give their opinions about the use of the dictionary that they consulted.
Results of the experiment show that there are no significant differences related to the number of correct answers of the DAELE groups with

respect to the control groups. We find, however, qualitative differences with regard to what students miss or value in every dictionary. The test confirms that the current approach taken in the preparation of DAELE, in which we aim to offer users the possibility of expanding or reducing the amount of information they see in response to a search, and to give them grammatical indications that are easier to understand and better suited to their needs.

> **An innovative medical learner's dictionary translated by means of speech recognition**

EERENBEEMT, ARNOUD VAN DEN

2 – The Dictionary-Making Process

I will discuss a medical dictionary based on the Keyword in context (KWiC) concept and speech recognition as a valuable tool. My daily work consists of compiling medical dictionaries for students and professionals (creating and updating complex and dynamic data) and creating medical spellcheckers.

Non-Anglophone medical students and health care professionals around the globe need an active command of professional English for their career. Yet the lexical tools available for acquiring these skills are few and insufficient: American and British explanatory dictionaries expect readers to be native speakers, while bilingual medical dictionaries are basically glossaries and provide unlabelled translations.

The only medical learner's dictionary in the world to date is the excellent Fachwortschatz Medizin by Michael and Ingrid Friedbichler, teaching English for medical purposes (EMP) in Austria. Their opus magnum, which helps non-native speakers to acquire language skills step by step, is structured using modular medical concepts and combines various lexical features:

- a monolingual dictionary: 100,000 medical terms grouped into 1400 sections with key headwords defined in simple English; contextualized with collocations and sample sentences demonstrating correct use, extracted from a 20-million-word corpus of medically authoritative texts;
- a semi-bilingual dictionary: support in the user's native language (German, Dutch) in the form of 42,000 translated keywords;
- a thesaurus: synonyms, antonyms and related terms;
- a domain-specific glossary: readers from all medical fields can focus on content relevant to their specialization;
- Windows edition: full-context search, customizable display (pronunciation, definition, translations, collocations), cross-references etc.

After acquiring the Dutch rights I realised that farming out the translation work would require me to extensively monitor the translators, who were discouraged by the highly condensed lexical content. I decided to translate the 42,000 indexed medical terms myself instead, using speech recognition and a 24" HD monitor to display my database content, a web browser, a word processor and two medical dictionaries. I developed voice-driven macros for automating 600,000 Google searches, creating 2000 records, searching dictionaries and my 52,000-record medical database etc. This allowed me to translate up to 400 terms per day.

www.pinkhof.nl/medisch-engels: full Windows edition 10-day trial period, free download of 60-page sample PDF

> **Thinking out of the box – perspectives on the use of lexicographic text boxes**

GOUWS, RUFUS H. AND PRINSLOO, DANIE J.

2 – *The Dictionary-Making Process*

Although text boxes have become a common phenomenon in dictionaries relatively little attention has been paid to their presentation and to the motivation for their use and the type of data to be included in a dictionary in this specific way. Text boxes are salient dictionary entries and as such they are used to place more than the default focus on a specific data item. Dictionaries offer a variety of data types in text boxes such as guidance in terms of sense, contrasting related words, restrictions on the range of application, register, pronunciation, et cetera. The default presentation seems to be as article-internal microstructural entries within a typical relation of lemmatic addressing. Whereas some text boxes present data relevant to only the specific article, other text boxes, i.e. those with a synoptic assignment, also have relevance for other articles, namely a hybrid addressing relation, presenting both immediate and distant addressing. As devices employed in an extended compulsory microstructure care should be taken that text boxes do not become part of the compulsory microstructure and in so doing lose their significance and decrease the emphasis on the data included in the text boxes. The added value of text boxes may never be undermined by an over exposure of this device. Using both micro- and macrostructural text boxes offers exciting possibilities. Where dictionaries have a text production function data could be included in a text box to emphasise the use or non-use of certain combinations and collocations as well as proscriptive guidance. Of real importance is that lexicographers should realise that text boxes

are lexicographic devices that can really enhance the data transfer in dictionaries. Lexicographers should think out of the box and they should get out of the box of tradition and employ text boxes in bold, innovative and functional ways.

> Guiding principles for the elaboration of an English-Spanish dictionary of multi-word expressions

GREGORIO-GODEO, EDUARDO DE

2 – *The Dictionary-Making Process*

So-called *word combinations* – also referred to as *multi-word combinations* or *multi-word expressions* – take shape when certain words regularly combine with certain other words or grammatical constructions. When exploring the word combinations of a language, both collocations and idiomatic expressions to a large extent examined. Collocations and idioms are usually taken to be multi-word expressions whose meaning is more than the sum of the meaning of their components.

Focusing on multi-word expressions in bilingual dictionaries, this contribution accounts for an ongoing research and editorial project guiding the elaboration of an English-Spanish dictionary of multi-word combinations. After presenting the lexicographic process leading the elaboration of the dictionary as such, this contribution will proceed to describe the principles determining the inclusion of entries and their presentation in the dictionary.

The rationale for this project is based on current lexicographic practices (Hartman 2001) having comprised four stages: (1) pre-lexicographic work, which consisted of a thorough examination of the market of English-Spanish dictionaries given the lack of specific dictionaries dealing with multi-word expressions in this area; (2) the research undertaken for the elaboration of the macrostructure of the dictionary and the use of various sources (e.g. existing English monolingual or multi-word dictionaries; bilingual dictionaries, and corpora), especially as far as usage examples, equivalents and their idiomaticity is concerned; (3) description issues, with a special emphasis on both the description of the multi-word expressions included in the dictionary, and the actual structure of dictionary entries; and (4) final formatting, which entails final presentation and revision prior to editing and publishing the dictionary.

Considering Spanish-speaking students of EFL and – to a lesser extent – translators as the potential users of this dictionary, this contribution will conclude with some final remarks of the educational implications of the project herein presented.

> **La segunda y tercera ediciones del *Diccionario Básico Escolar***

MIYARES BERMÚDEZ, ELOÍNA; ARTOLA ZUBILLAGA, XABIER; ALEGRÍA LOINAZ, IÑAKI; ARREGI IPARRAGIRRE, XABIER; RUIZ MIYARES, LEONEL; ÁLAMO SUÁREZ, CRISTINA AND PÉREZ MARQUÉS, CELIA

2 – *The Dictionary-Making Process*

En julio del 2003 se publicó la primera edición del *Diccionario Básico Escolar* (DBE), obra desarrollada en el Centro de Lingüística Aplicada de Santiago de Cuba y orientada a un mejor dominio del idioma español por parte de sus destinatarios: estudiantes del segundo ciclo de primaria (5to y 6to grados), secundaria básica y preuniversitario.
Gracias a la inestimable colaboración del Grupo IXA de la Universidad del País Vasco y al Instituto Cubano del Libro de Cuba, se presentó la posibilidad de realizar la segunda y tercera ediciones del DBE, por lo que nuestro grupo lexicográfico emprendió la laboriosa y complicada tarea de mejorar y arreglar algunas entradas, además de agregar nuevos artículos a esta importante obra de consulta.
El *Diccionario Básico Escolar* está disponible en tres soportes: impreso, en CD y en INTERNET (http://ixa2.si.ehu.es/dbe/index.html) y la segunda y tercera ediciones del mismo incluyó la completa revisión de sus tres soportes.
Los diccionarios son 'organismos vivos'; un diccionario que posea varias ediciones tiene que revisarse constantemente, pues siempre habrá entradas que mejorar, otras que añadir y corregir los errores humanos, hasta llegar a una obra casi perfecta.
En este trabajo pretendemos analizar el entorno de edición de diccionarios *leXkit*, las características de la segunda y tercera ediciones del *Diccionario Básico Escolar,* sus resultados y una comparación con la primera edición, donde se demuestra la 'vitalidad' de esta herramienta lingüístico-pedagógica.

> **The Living Lexicon: Methodology to set up Synchronic Dictionaries**

NAZAR, ROGELIO AND AZARIAN, JENNY

2 – *The Dictionary-Making Process*

In this paper, we want to investigate the subset of the vocabulary of a given language or dialectal variant which is in actual use in the discourse of a linguistic community in order to set up a synchronic dictionary. The aim of this article is, thus, to develop a methodology for acquiring the nomenclature of synchronic dictionaries in a systematic way. To do this, we consider two kinds of operations: addition of entries –the birth of words, or Neology- and removal -the death of words, Desuetude, as we

call it here. The methodology consists in contrasting dictionaries of a language (or dialectal variant) to find the intersection of the vocabulary, and to compare the vocabulary of the dictionaries with the vocabulary of a diachronic corpus. Such a methodology enables us to answer the following research questions: 1) what proportion of the vocabulary is shared by most dictionaries, 2) what proportion of units of each dictionary is no longer in use and 3) what proportion of the vocabulary units in use today is still not registered in the dictionaries. These three questions are central to the definition of the ideal headword. In a pilot experiment in Peninsular Spanish, we combine the study of the main dictionaries of this language variant with diachronic studies using corpus statistics on Spanish newspaper archives.

> **Lingvo Universal English-Russian Dictionary: Making a Printed Dictionary from an Electronic One**

ANOKHINA, JULIA

3 – Reports on Lexicographical and Lexicological Projects

Lingvo Universal English-Russian Dictionary (*Lingvo UERD*) was the first electronic English-Russian dictionary published in Russia. It appeared in 1990, as part of the *Lingvo* software produced by the company *ABBYY*. Unlike many other dictionaries available on *Lingvo*, which are licensed electronic versions of high-quality paper editions, *Lingvo UERD* is the fruit of the company's own lexicographic research. As the dictionary database grew further, it was transformed into a multifunctional database, used to produce different kinds of dictionaries. The first printed edition based on its content was the *ABBYY Lingvo Comprehensive English-Russian Dictionary*, published in 2007. It was designed as an English-Russian dictionary for professional users and advanced learners; the dictionary entries were edited in the in-house DWS *ABBYY Lingvo Content*. The 2nd revised edition of it was initiated in 2008 by the publishing house *ABBYY Press*; the current article reports on this project.

While preparing the dictionary lexicographers faced a whole range of problems related to different access to electronic vs. printed dictionary data, to different user tasks while accessing them and to the specific character of the *Lingvo* software format. Many difficulties were solved by *ABBYY* programmers who adjusted export algorithms of the DWS and improved its interface. All those improvements were added to the latest version of the DWS *ABBYY Lingvo Content*.

Present-day dictionary databases tend to include as much linguistic data as possible in order to be used as a basis for different kinds of dictionaries, including printed editions. As an electronic database is a

big hypertext comprising multiple links and different kinds of specific data which cannot be exported to the 'paper' format, making a paper dictionary from such a database may be quite a challenging task. Working hand in hand with the publishing house editors enabled us to minimize the inevitable losses resulting from such a procedure. The other result of this work was the creation of a printed dictionary more in line with the needs of modern users, presented in a more convenient and user-friendly way.

> Database of ANalysed Texts of English (DANTE): the NEID database project

ATKINS, B.T. SUE; KILGARRIFF, ADAM AND RUNDELL, MICHAEL

3 – Reports on Lexicographical and Lexicological Projects

DANTE is a lexicographic project where the end product is not a dictionary but a lexical database resulting from in-depth analysis of corpus data. The users of DANTE are not the dictionary-using public but the lexicographic teams who will develop dictionaries and computer lexicons from it. This project is the source-language analysis stage of the New English-Irish Dictionary (NEID: http://www.forasnagaeilge.ie/), being developed for Foras na Gaeilge, Dublin (FnaG: http://www.forasnagaeilge.ie/). The project was designed and carried out by the Lexicography MasterClass (http://www.lexmasterclass.com). The database covers approximately 50,000 headwords and 45,000 compounds, idioms and phrasal verbs, using over 40 datatypes in their lexical description. It was created in the course of 2.5 years by LexMC's 15-strong lexicographic team, managed by Valerie Grundy, Managing Editor; the project administration is in the hands of Diana Rawlinson, Project Administrator.

What makes DANTE special is the application of an existing methodology across the whole lexicon, extremely systematically and at an unprecedented level of detail. Amongst other aspects of this project, we describe:

- improving the reliability of schedule and workflow by classifying, before the compiling started, over 50,000 headwords according to type and complexity;
- the systematic use of 68 model 'template' entries;
- a new approach to quality control, combining conventional entry-editing by senior team members with the use of complex search scripts that list all entities of a specific type and allow rapid checking for accuracy;
- the customisation of the Sketch Engine (http://www.sketchengine.co.uk/) corpus query software, with a corpus of 1.7bn words;
- the use of IDM's Dictionary Production System (DPS: http://www.idm.fr/products/dictionary_writing_system/27/)..

The DANTE database is a rare, possibly unique, beast: a rich and comprehensive lexicographic analysis on linguistic principles, prepared on a substantial budget by a large team of professional lexicographers, and uncompromised by the needs of accessibility to non-linguist users.

> **Quo Vadis Lexicography at the Institute for Dutch Lexicology?**

BEEKEN, JEANNINE

3 – Reports on Lexicographical and Lexicological Projects

In this paper, we will first introduce the Institute for Dutch Lexicology. We will present an overview of the INL-dictionaries online, being the Dictionaries of Old Dutch (ca. 475 – 1200), Early Middle Dutch (1200 – 1300), Middle Dutch (1250 – 1550), the Dictionary of the Dutch Language (WNT, 1500 – 1970s), the Etymological Dictionary of Dutch, the General Dutch Dictionary (ANW, 1970s till 2015). Thirdly, we will present the Language Bank (Taalbank Nederlands) and its main tasks. Finally, we will elaborate on three U-turns, namely a first U-turn: from manual labour and printed material to computational linguistics and the internet, a second U-turn: from single functionality to multiple functionality, a third U-turn: from stand-alone product to spin-offs, linking and integration. We will finish with some thoughts and ideas answering the following question: quo vadis lexicography at the INL?

> **Time to say goodbye?**
> **On the exclusion of solid compounds from the Swedish Academy Glossary (SAOL)**

BERG, STURE; HOLMER, LOUISE AND SKÖLDBERG, EMMA

3 – Reports on Lexicographical and Lexicological Projects

The Swedish Academy Glossary, SAOL (short for *Svenska Akademiens ordlista*) is a monolingual glossary, first published in 1874. The latest edition, SAOL13, was published in 2006 and the next edition, SAOL14, is planned for 2015.

This article concerns the revision of the lemma list in SAOL, with special focus on the exclusion of transparent solid compounds. There are about 88,000 solid compounds in the 13th edition of the Glossary, i.e. 70 % of the total number of lemmas (125,000). Since there are almost infinite possibilities of creating new words in Swedish, the printed Glossary obviously only includes a sample of the contemporary Swedish vocabulary.

With improved lexicographic tools and an enlarged text corpus, the editors of SAOL14 have great possibilities of making more accurate decisions

when including new solid compounds and excluding others from the lemma list. The discussion is above all based on the solid compounds including the noun *kalkyl* ('calculation', 'estimate', 'calculus').

> **FACKELLEX – Zur Struktur des Schimpfwörterbuches**

BREITENEDER, EVELYN

3 – Reports on Lexicographical and Lexicological Projects

In 2008, the work on the 'Schimpfwörterbuch' (Dictionary of Insults and Invectives), the second part of the Fackellex dictionary, was brought to an end. FACKELLEX is a so called dictionary in the field of textlexicography. The 'Schimpfwörterbuch' was compiled in the planned tripartite structure, developing the dictionary volumes named ALPHA, CHRONO and EXPLICA. The paper will show different methods of presenting text information in a dictionary.

During the work on the dictionary, some 200.000 invective expressions were identified on the 22.586 pages of the 'Fackel' ('The Torch') edited and written by Karl Kraus, 2775 of which were selected to be represented as keywords in the ALPHA volume – the alphabetic list of the dictionary. Selection was performed according to linguistic and semantic criteria, keywords were furnished with short excerpts from the original text. Main tasks during this phase of the project were the constitution of a list of candidates and the presentation of the extensive material. The focus in the work on the ALPHA section was the description of invective terms, particular constructions and examples of Karl Kraus's creativity in coining new words making use of text lexicographic methods. ALPHA is made accessible through three different indexes which were created making use of up-to-date IT technology as part of a cooperation within the 'Centre for Cultural Research' between the departments of AAC and Fackellex.

CHRONO – the chronological list of the 'Schimpfwörterbuch' – displays in chronological order roughly a fifth of the data contained in ALPHA and offers the reader a larger context. This part of the dictionary is designed to pursue the development of the author's creativity in coining and using words within the ›Fackel‹ as a whole and to display these phenomena in the context of a particular page of the journal.

EXPLICA – short for explanatory notes – is to fulfill two requirements: it contains the dictionary editor's explanatory texts on which the project was based, which were written as part of the seperate 'PARATEXTE' project. In addition, it contains ›Wichtiges von Wichten‹, the last article of the ›Fackel‹ which can be seen as the primary source of inspiration for this ›Schimpfwörterbuch‹. Passages of this text that were identified as invectives where highlighted and commented upon in a selective way.

> **Improving the representation of word-formation in multilingual lexicographic tools: the MuLeXFoR database**

CARTONI, BRUNO AND LEFER, MARIE-AUDE

3 – Reports on Lexicographical and Lexicological Projects

This paper introduces a new lexicographic resource, MuLeXFoR, which aims to present word-formation processes in a multilingual database designed for both language specialists (e.g. linguists, terminologists, lexicographers, NLP specialists) as well as second-language (L2) learners and trainee translators. Morphological items (e.g. affixes, compound parts, combining forms) and processes (prefixation, suffixation, compounding, conversion, etc.) pose major challenges for lexicographic work, especially with respect to the design of bilingual and multilingual resources. It is well-known that derivational affixes can take part in several word-formation rules and that, conversely, rules can be realised by means of a variety of affixes. In view of this complexity, it is often difficult to (1) provide enough information to help users understand the meaning(s) of an affix and the (near-)synonymy relations between affixes and (2) become familiar with the most frequent strategies used to translate the meaning(s) conveyed by these affixes. In fact, traditional dictionaries often fail to achieve this goal. The MuLeXFoR database tries to take advantage of recent advances in morphological description and the development of electronic multi-access database systems. The database relies on the lexematic approach to word-formation, which is especially helpful to represent morphological processes cross-linguistically. In addition, it has been entirely implemented in a multi-access database interface. The prototype described in this paper so far centres around prefixation in English, French and Italian. Two interfaces are currently available: a comprehensive interface aimed at morphological and lexicographic investigations by language specialists (*MuLeXFoR-Linguists*) and a second interface designed for second-language learners or trainee translators (*MuLeXFoR-Learners*).

> **Author Dictionaries Revisited: Dictionary of Bohumil Hrabal**

ČERMÁK, FRANTIŠEK AND CVRČEK, VÁCLAV

3 – Reports on Lexicographical and Lexicological Projects

With a view to continue the line of author dictionaries, started by that devoted to Karel Čapek (2007), a second dictionary, basically following the first, has been compiled, namely that of Bohumil Hrabal (2009), an influential and major figure of the contemporary literary scene. The idea to have more of comparable and corpus-based dictionaries of this

type that would ultimately enable comparison and through the prism of some of the best masters of the language to view the Czech language in development, has been made possible only recently, with the existence of corpora and thanks to techniques developed by corpus linguistics. A number of new lexicographic and computational features, never used before (with the exception of K. Čapek's dictionary), have been tried verifying options how to best put into practice general theoretical ideas, such as when finding best collocations that could be included in the dictionary.

> **The Faroese-Italian Dictionary - An attempt to convey linguistic information concerning the Faroese language as well as information about the culture of the Faroe Islands**

CONTRI, GIANFRANCO

3 – Reports on Lexicographical and Lexicological Projects

In 2004 Føroya Fróðskaparfelag, the Academy of the Faroe Islands, published the *Dizionario Faroese-Italiano / Føroysk-Italsk orðabók*, the first bilingual dictionary of the Faroese and Italian languages. The dictionary has 632 pages and includes 14.850 headwords, plus a few hundred sub-headwords within the relevant single entry. It was Professor Jørgen Stender Clausen of Pisa University in Italy who suggested that I compile this dictionary, and I carried it out working at the Department of Faroese Language and Literature of the University of the Faroe Islands. The dictionary is the result of some years' work and of the indispensable advice I was given by the lexicographical consultant Jógvan í Lon Jacobsen, and of the help of several Faroese and Italian collaborators. The dictionary is an attempt to create a practical means to help an Italian-speaking visitor or student to make acquaintance with both the Faroese language and the culture of the Islands, and also useful for Faroese meeting Italian-speakers on their travels. The differences between the two languages (with some structural features showing diversity), and the cultural differences between the linguistic areas (the respective everyday vocabulary is different), had to be dealt lexicographically. The compilation of the dictionary has not followed any existing lexicographic model: the list of headwords, the structure of the entries and the graphics are the result of research and experiment. One result, among others, is that many terms are related to the needs of a visitor or student interested not only in the language of the Faroe Islands but also their culture, and that therefore some entries are a combination of linguistic and 'cultural-encyclopedic' information.

> **Covering All Bases: Regional Marking of Material in the New English-Irish Dictionary**

CONVERY, CATHAL; Ó MIANÁIN, PÁDRAIG AND Ó RAGHALLAIGH, MUIRIS

3 – Reports on Lexicographical and Lexicological Projects

The New English-Irish Dictionary is a government-sponsored project that began in 2000 and is due for completion in 2012. The aim is to produce a modern bilingual dictionary containing c. 40,000 headwords which is to be published in both printed and electronic formats. When published this dictionary will be the first major dictionary published for Irish in over 40 years. The project is currently at the translation phase, and this paper focuses on the approach taken to attempt cover dialect variations in the modern spoken language. The methodology employed was divide the headword list into three distinct categories, each requiring a different level of translation.
Given the time and budgetary constraints of the project it was decided that only the 1000 (approx) most frequently occurring lemmas could receive a full dialectal profile. Translators from each of the three main dialects translate each entry, passing the entry on to a translator from the next dialect as they complete their part of the process. This translation work is carried out without reference to written sources. Once a translator from each main dialect has completed their work the entry is checked for completeness against set sources and labelled accordingly.
The main advantages of this process are as follows.

- It captures current translations that may not be covered in existing outdated sources.
- It provides a dialectal profile of words, phrases and usages.
- It enables an element of dialectal marking in the final product, particularly in the electronic version.
- It enables the option to customise the electronic version, fronting any particular dialect.
- A given dialect may be selected as the default pronunciation in the electronic version.
- It enriches the bilingual database creating a useful research resource for other academic research projects.

> **Software Demonstration of the Dictionary of the Flemish Dialects and the pilot project Dictionary of the Dutch Dialects**

DE TIER, VERONIQUE AND VAN KEYMEULEN, JACQUES

3 – Reports on Lexicographical and Lexicological Projects

A. Dictionary of the Flemish Dialects

The relational database of the *Dictionary of Flemish Dialects* works under

Oracle. The WVD input database (*bronsoorten* = sources) consists of subdatabases of one or more questionnaires. Once all the data have been put into the correct subdatabases, the lexicographer can start compiling the dictionary by selecting the concepts that are to be put in one particular fascicle, e.g. all the selected concepts for 'sheep'. This is done in a new database structure, called 'publications'. After selecting the data from the different sources, the lexicographer automatically can generate a dictionary article with all the results for one concept. From this point onwards, the dictionary article can be compiled and lexicological decisions have to be made. This database also forms the basis for drawing the word maps in MapInfo and for making a text file ('Wetenschappelijk Apparaat' (Scientific Database)) in which every entry with the different lexemes for that particular concept is followed by codes that indicate the location of the words.

B. Pilot project : the *Dictionary of the Dutch dialects*

The software of the *Dictionary of Dutch Dialects* works under the oracle platform as well. After digitizing dialect dictionaries by scanning and ocring and after correcting the Word files of these dictionaries, the headwords are put into bold and two Hard Returns are inserted after each dictionary article. This Word document, converted into a standard XML file, can be imported into the database through the application built for this database. For each dictionary it is necessary to write a new custom script, which generates the XML-file by means of typographical conventions. Once the XML-files are uploaded, the database of the *Dictionary of the Dutch Dialects* can be made. The editors then may enrich the database with dutchification, translation and markers. The next step is to connect this database to a website with search facilities.

> **The Style Manual for Monolingual Lụgbarati Dictionary**

DRAMANI, SAIDI

3 – Reports on Lexicographical and Lexicological Projects

To compile a monolingual general-purpose Lụgbarati dictionary, a Style Manual based on the format of Makerere Institute of Languages was developed (Kiingi, 2004). It was the blue print for the process of compilation. Lụgbarati terminology for linguistics has hitherto been lacking. Words were coined using functions of the word classes. The coinages were used to give ancillary information on the lexical items being defined. The research involved developing a style manual, compiling the dictionary, testing it for acceptability, and analysing the testing outcomes. The corpus used was a

198-page list of vocabulary in Crazzolara's book; *A Study of Lugbara (Ma'di) Language* (1960:175-373), and a 25-page list of words in Dalfovo's collection of Lụgbara proverbs; *Lugbara* (sic) *Proverbs* (1984:249-274).

> **An inverted loanword dictionary of German loanwords in the languages of the South Pacific**

ENGELBERG, STEFAN

3 – Reports on Lexicographical and Lexicological Projects

The paper reports on a dictionary of German loanwords in the languages of the South Pacific that is compiled at the Institut für Deutsche Sprache in Mannheim. The loanwords described in this dictionary mainly result from language contact between 1884 and 1914, when the German empire was in possession of large areas of the South Pacific where overall more than 700 indigenous languages were spoken.
The dictionary is designed as an electronic XML-based resource from which an internet dictionary and a printed dictionary can be derived. Its printed version is intended as an 'inverted loanword dictionary', that is, a dictionary that – in contrast to the usual praxis in loanword lexicography – lemmatizes the words of a source language that have been borrowed by other languages. Each of the loanwords will be described with respect to its form and meaning and the contact situation in which it was borrowed. Among the outer texts of the dictionary are (i) a list of all sources with bibliographic and archival information, (ii) a commentary on each source, (iii) a short history of the language contact with German for each target language, and perhaps (iv) facsimiles of source texts.
The dictionary is supposed to (i) help to reconstruct the history of language contact of the source language, (ii) provide evidence for the cultural contact between the populations speaking the source and the target languages, (iii) enable linguistic theories about the systematic changes of the semantic, morphosyntactic, or phonological lexical properties of the source language when its words are borrowed into genetically and typologically different languages, and (iv) establish a thoroughly described case for testing typological theories of borrowing.

> **The development of scholarly lexicography of the Estonian Language as a Second Language in an historical and a theoretical perspective**

KALLAS, JELENA

3 – Reports on Lexicographical and Lexicological Projects

This paper aims to provide an overview of the development of scholarly lexicography of the Estonian language as a second language in an

historical and a theoretical perspective. The paper describes what kind of information is presented traditionally in dictionary entries on the level of morphology, derivation, syntagmatic relationships and paradigmatic relationships. In addition, taking into consideration theoretical and practical viewpoints of modern lexicography on what kind of information should be presented in a dictionary entry so that the dictionary could be classified as a production dictionary (Apresjan (ed.) 2006; Atkins & Rundell 2008; Bo Svensén 2009; Novikov 2001; Siepmann 2006), the author is going to illustrate what kind of information should be added into the entries of a learners' dictionary of the Estonian language as a second language so that they could be used as production dictionaries.
In an historical perspective the analysis of the learners' dictionaries, which were published during the last 160 years, indicated that dictionary compilers provide dictionary users mostly with information about inflectional formation; meanwhile, the information about word formation (derivatives, compounds), syntagmatic and paradigmatic relationships is almost neglected. On the other hand, learners' dictionaries meant for speakers of Estonian as a first language provide much more information: the information about inflectional formation, word formation, synonyms, antonyms, paronyms is presented explicitly. The information about syntagmatic relationships is presented mostly implicitly by means of examples at the level of phrases, clauses and sentences.
The author puts forward detailed proposals for what kind of formal (inflectional formation, derivatives, compounds), semantic (mostly content-paradigmatic information) and syntagmatic (syntactic valency, collocations, idioms) characteristics should be given in a dictionary of the Estonian language as a second language and demonstrates practical implementations of explicit systematic description of syntactic valency and collocations of different parts of speech (nouns, adjectives, adverbs, verbs, quantifiers).

> WikiProverbs – Online Encyclopedia of Proverbs

KATS, PAVEL

3 – Reports on Lexicographical and Lexicological Projects

The WikiProverbs project was envisioned as a free online multilingual dictionary of proverbs, edited by the community. The idea behind the project was to address the difficulty of translating proverbs across the languages and to create a public repository of multilingual equivalents of proverbs that will serve language professionals, such as: writers, translators, journalists, as well as language enthusiasts. From its inception the project was conceived as a non-profit humanitarian enterprise for the sake of Internet users.

> **Stichwort, Stichwortliste und Eigennamen in elexiko: Einflüsse der Korpusbasiertheit und Hypermedialität auf die lexikografische Konzeption**

KLOSA, ANNETTE; SCHNÖRCH, ULRICH AND SCHOOLAERT, SABINE

3 – Reports on Lexicographical and Lexicological Projects

Die Überschrift des Beitrags impliziert dessen Gliederung in zwei größere thematische Abschnitte: der erste, allgemeine widmet sich Überlegungen zu Stichwort und Stichwortliste, der zweite, speziellere erörtert die Behandlung von Eigennamen in *elexiko*.
elexiko (www.elexiko.de) ist ein am Institut für Deutsche Sprache in Mannheim entstehendes Online-Wörterbuch zur deutschen Sprache. Die methodische Basis für die redaktionelle, lexikografische Erarbeitung von Wortartikeln ist das Prinzip der Korpusbasiertheit. Voraussetzung für dessen methodische Umsetzung ist, dass für jedes Stichwort (und seine Lesarten) Belege in ausreichender Anzahl und Qualität im *elexiko*-Korpus vorhanden sind. Um das zu gewährleisten wurde auch die Stichwortliste komplett neu erstellt, und zwar auf der Basis von Korpora des geschriebenen Deutsch seit 1946. Im ersten Teil des Beitrags werden grundsätzliche Gedanken zur Erarbeitung einer adäquaten Stichwortkonzeption im Rahmen eines Online-Wörterbuches dargelegt, Sonderfälle und Ausnahmen vorgestellt sowie die Vorgehensweise bei der korpusbasierten Erstellung der *elexiko*-Stichwortliste skizziert.
Ausgehend von der Definition von Eigennamen erörtert der zweite Teil des Beitrags die gängige lexikografische Behandlung von Eigennamen in allgemeinsprachigen Wörterbüchern und stellt Überlegungen dazu an, wie Eigennamen in Abgrenzung zu Gattungsbezeichungen lemmatisiert werden sollten. Dabei stellt sich für ein Online-Wörterbuch wie *elexiko*, dessen Schwerpunkt der lexikografischen Beschreibung auf der Bedeutung und Verwendung von Stichwörtern liegt, die Frage, in welcher Form die lexikografische Behandlung von Eigennamen erfolgen soll. Außerdem thematisiert dieser Beitrag die Behandlung von Eigennamen in *elexiko* hinsichtlich ihrer Erfassung, Klassifizierung und Darstellung und erläutert unterschiedliche Angabetypen. Ein Ausblick auf Suchoptionen zu den Eigennamen schließt die Überlegungen ab.

> **Orthographical Dictionaries: How Much Can You Expect? The Danish Spelling Dictionary Revis(it)ed**

LORENTZEN, HENRIK

3 – Reports on Lexicographical and Lexicological Projects

Orthographical dictionaries constitute a particular and rather specialised subclass of dictionaries. This contribution offers a presentation of the

ongoing revision of a spelling dictionary (for Danish) and a discussion of some of the general and specific issues that have arisen during the project. Firstly, the historical background is described, a brief overview of the many editorial changes is provided, and lemma selection, variant forms and definitions are discussed in some detail. Particular interest is paid to the number and character of the included headwords, to the problems of (too many) variant forms and to the difficulties involved in providing definitions in a dictionary whose main purpose is to inform about correct spelling. Secondly, the field of official and unofficial orthographical dictionaries in Denmark is compared to that of some other countries of northern Europe: Sweden and Germany, and it is shown how the forthcoming edition of the Danish spelling dictionary is inspired by the other dictionaries. Finally, the conclusion engages in a discussion of the necessity of this particular type of dictionary, which to this author seems somewhat questionable.

> **A language on the back foot: The Afrikaans lexicographer's dilemma**

LUTHER , JANA

3 – Reports on Lexicographical and Lexicological Projects

Afrikaans originated in the variants of Dutch that developed at the southern tip of Africa during the 17th and 18th centuries. In the 19th century, when English began to overtake Dutch as the high-function language in the Cape, proponents of Dutch and Afrikaans put up a resistance, and during the 20th century the functions of Afrikaans expanded until it could take its place alongside Dutch and later stand with equal status next to English. As an official language Afrikaans reached back to Dutch a second time to develop into a full-fledged language. But its heyday could not last indefinitely. In recent decades the milieu of Afrikaans speakers has changed radically. Political upheaval, technological advances, new areas of specialisation, the lightning pace of new developments have thrust Afrikaansers into the thick of the world-wide explosion of knowledge which demands efficient communication. A third reversion to Dutch is out of the question. The path between Afrikaans and Dutch has become overgrown; few present-day users of Afrikaans still walk along it. Likewise, to the average Dutch man and woman, Afrikaans today is a distant language. In the multilingual South Africa, where English dominates, the effect of the contact with English on Afrikaans is undeniable. A serious threat to Afrikaans is its loss of status in the judiciary, the administration, education and as a scientific language. Against this backdrop the *Handwoordeboek van die Afrikaanse Taal* (*HAT*) – a household name among Afrikaans speakers, comparable to the Dutch

'Dikke van Dale' – is subjected to scrutiny: After its 'golden age', how well has the *HAT* kept pace with Standard Afrikaans in transition? Can it keep in step with the unstoppable, irreversible changes of the time and in the language today? Or will Afrikaans's flagship dictionary, in a decade or so, lose its relevance for the Afrikaans user?

> **Phonetic Transcriptions for the New Dictionary of Italian Anglicisms**

MAIRANO, PAOLO

3 – Reports on Lexicographical and Lexicological Projects

This paper describes the work that has been done concerning the phonetic transcriptions for the *New Dictionary of Italian Anglicisms* directed by Prof. Pulcini (University of Turin): the dictionary contains both transcriptions of how Italians pronounce anglicisms and of how the corresponding English words are pronounced by native speakers of English. We shall explain how different pronunciation variants were selected for inclusion in the dictionary and how the transcriptions of anglicisms had to be adapted to the phonology and phonetics of Italian. A discussion will follow about the effects caused by the juxtaposition of English and Italian transcriptions. In fact, because of the intereference of the two phonetic and phonological systems, traditional conventions were in some cases abandoned in favour of more accurate phonetic transcriptions: this has been done with the aim of illustrating the most remarkable differences between the pronunciation of the words by Italian and English speakers.

> **Centre for Bilingual Lexicography at Tbilisi State University, Georgia. Projects, Methods, History**

MARGALITADZE, TINATIN

3 – Reports on Lexicographical and Lexicological Projects

The first bilingual dictionary of the Georgian language, Georgian-Italian was compiled and published in 1629 in Rome. Between 1629 and 1870 approximately ten European-Georgian dictionaries were compiled.
At the beginning of the 19th century Georgia became a part of the Russian Empire. Since that time the major emphasis has been placed on Russian-Georgian lexicography. As a result of such an approach, bilingual lexicography of the Georgian language suffered in respect to European languages.
Even when the first English-Georgian, or other European-Georgian dictionaries appeared from the 1940s, they were mere translations of European-Russian dictionaries, which led to numerous inaccuracies and even gross mistakes.
The same erroneous lexicographical principles became the basis of

compilation of A Comprehensive English-Georgian Dictionary (CEGD), initiated by the Department of English Philology of Tbilisi State University back in the 1960s. The decision was made to translate the New English-Russian Dictionary, edited by I. Galperin.
After examining the existing material, the Editorial Staff of CEGD (established in the 1980s) arrived at the conclusion that it was impossible to edit the material in the form in which it was executed. The Editorial Staff developed entirely new principles for the creation of CEGD.
The method of the analysis of definitions of English Dictionaries was identified as the basic technique for the investigation of the semantic structure of English lexical units.
The process of revision and editing of the material of CEGD has continued for 25 years.
The publication of CEGD started in 1995 in fascicles on a letter-by-letter basis. By now, thirteen fascicles of CEGD have been published, from letters A to O.
The Internet version of CEGD was launched in February 2010.
Other projects of the Lexicographic Centre include: 'English-Georgian Learner's Dictionary';
'English-Georgian Military Dictionary' (first publication of the series of specialised English-Georgian Dictionaries).

> **Crossing borders in lexicography:**
How to treat lexical variance between countries that use the same language

PARQUI, JAAP; BOON, TON DEN AND HENDRICKX, RUUD

3 – Reports on Lexicographical and Lexicological Projects

In the past decades the identity of Belgian Dutch has changed considerably. It no longer tries to copy Netherlands Dutch, but is following its own course. This development should be reflected in Dutch dictionaries. For different dictionaries (e.g. bilingual and explanatory dictionaries) and for different categories of words (e.g. juridical or informal words) different strategies should be adopted.

> **Better Nicely Linked than Poorly Copied.**
Historical and Regional Dictionaries of Dutch Digitally United

SCHOONHEIM, TANNEKE AND DE TIER, VERONIQUE

3 – Reports on Lexicographical and Lexicological Projects

The *Woordenboek der Nederlandsche Taal* (Dictionary of the Dutch Language, WNT) has been freely available online since January 2007 (http://wnt.inl.nl). Compared to the original (printed) dictionary, the search facilities

have been considerably expanded. For instance, you can now search for a headword using modern spelling, and submeanings and citations can be displayed or omitted on demand.
Another innovation is that headwords in the WNT are now linked to external information, for example, to language maps, pictures and etymological information. At the moment, we add links to the available dialect material, starting with the large dictionaries of the dialects of Flanders, Brabant and Limburg. In this contribution we describe how this is done.

> **Dutch Lexicography in Progress: the *Algemeen Nederlands Woordenboek* (ANW)**

SCHOONHEIM, TANNEKE AND TEMPELAARS, ROB

3 – Reports on Lexicographical and Lexicological Projects

The *Algemeen Nederlands Woordenboek* (ANW – Dictionary of Contemporary Dutch) is a project of the Institute for Dutch Lexicology in Leiden, the Netherlands. It is an online corpus-based, scholarly dictionary of contemporary standard Dutch in the Netherlands and in Flanders, the Dutch speaking part of Belgium. It describes the Dutch vocabulary from 1970 onwards.
The ANW is aimed at a large audience, ranging from professional linguists to students and puzzlers. It provides information on form, content and use of words belonging to the general vocabulary of Dutch. It has an elaborate structure which aims to simplify the retrievability of words and meanings for the user compared to existing digital dictionaries. The semagram plays an important role in this, but so do various other innovative elements in the structure.

> **Wurdboek fan de Fryske taal/Dictionary of the Frisian Language online: new possibilities, new opportunities**

SIJENS, HINDRIK AND DEPUYDT, KATRIEN

3 – Reports on Lexicographical and Lexicological Projects

The Wurdboek fan de Fryske Taal (Dictionary of the Frisian Language, WFT) describes the vocabulary of the Modern West Frisian language and consist of 25 volumes of 400 pages each. The dictionary contains more than 100,000 entries. This paper is intended to show that an electronic version of the WFT, once the data have been converted to state-of-the-art standards and made available to the public by means of an advanced retrieval application, will be a modern lexicographical resource of significant value.

Integrating the WFT into the dictionary component of the *Geïntegreerde Taalbank Nederlands* (Integrated Language Database of Dutch, GTB) of the *Instituut voor Nederlandse Lexicologie* is the obvious means to reach this goal. In order to create more ways of searching the dictionary entries, data accessibility has to be enhanced by explicit tagging of information categories which can be exploited by a retrieval application.
The process of implementing the online version of WFT took place in several stages: First the existing database had to be repaired and optimised. Mistakes and inconsistencies had to be repaired. The logical structure had to be parsed and tagged with XML mark-up. Furthermore the newly created XML database had to be enriched with TEI encoding. And, finally the dictionary was incorporated into the GTB application.
The WFT has been incorporated into the online dictionary application of the Dutch language bank, and so is freely available to a large audience, allowing interested parties to search in one of the most complete Frisian dictionaries, and to explore the Frisian language in relation to Dutch.

> **The principles and structure of the Estonian Etymological Dictionary**

SOOSAAR, SVEN-ERIK

3 – Reports on Lexicographical and Lexicological Projects

The Estonian Etymological Dictionary (EED) has been a project of the Institute of the Estonian Language (IEL) since 2003. Due to the urgent necessity for an etymological dictionary it was decided to start from a short and not too detailed version tailored for the general public with no philological background and to broaden this version later in order to compile a scientific dictionary. The next step involved concrete decisions about the material to be included into the first version of the dictionary.

> **Economicus: A New Conception of the Bilingual Business Dictionary**

STORCHEVOY, MAXIM A.

3 – Reports on Lexicographical and Lexicological Projects

The paper is devoted to the *Economicus* project – English-Russian Dictionaries in Economic, Management and Finance – which is built on the new conception of bilingual business dictionary with rich, reliable and user-friendly lexicographical information for ordinary users (students, managers, translators etc.) as well as for researchers of language. The Economicus uses a rather sophisticated and advanced concept of entry with multiple zones which relies heavily on advantages of electronic entry demonstration and especially on the possibility of hiding and showing zones and subzones at user discretion. The latter feature creates

enormous opportunities for the lexicographer to develop a rich but still user-friendly content of the entry.

The project is based on the alliance of economists and linguists. The linguistic expertise for the project was provided by the ABBYY software company who gave Economicus lexicographers access to a database specially designed for building dictionaries, its proprietary markup language and its linguistic corpus. ABBYY linguists took active part in developing the conception of Economicus entry and helping economic lexicographers to find a correct and effective approach to developing an up-to-date terminological dictionary.

The economic and business expertise for the project was provided by several dozens of professors of various educational institutions in Russia and abroad. The most important role was played by professors of Graduate School of Management (GSOM), St-Petersburg State University who took active part in evaluating and improving entries in corporate finance, management, marketing, international business and other business fields. In 2007 Graduate School of Management established the Translation and Lexicography Department where Economicus project has been developed since that time.

Economicus dictionaries are distributed with ABBYY Lingvo (as part of its basic dictionary collection and as additional downloads) and are accessible through a web-site http://dictionary.economicus.ru. The number of entries in Economicus is currently about 75 000.

> The ANW: an online Dutch Dictionary

TIBERIUS, CAROLE AND NIESTADT, JAN

3 – Reports on Lexicographical and Lexicological Projects

The *Algemeen Nederlands Woordenboek* (ANW) is a comprehensive online scholarly dictionary of contemporary standard Dutch in the Netherlands and in Flanders, the Dutch speaking part of Belgium (Moerdijk 2004, 2008; Moerdijk, Tiberius & Niestadt 2008). The dictionary focuses on written Dutch and covers the period from 1970 onwards. The dictionary was conceived as an online dictionary right from the outset and offers a range of search possibilities supporting both semasiological and onomasiological queries. A demo version of the dictionary[1] was launched at the end of 2009 (http://anw.inl.nl). This paper discusses the search application of the ANW dictionary. It focuses on the access strategies that are offered and on FunQy, the query language that was specifically developed for the project to facilitate implementation and future extensions to the search options offered by the ANW. Currently the demo version of the dictionary has just over 2000 registered users.

> **Pilot project: A Dictionary of the Dutch Dialects**

VAN KEYMEULEN, JACQUES AND DE TIER, VERONIQUE

3 – Reports on Lexicographical and Lexicological Projects

The lexicon of the traditional dialects in the Dutch language area is disappearing at a rapid pace. Three major regional dialect dictionaries, the *Dictionary of the Brabantic dialects* (WBD), the *Dictionary for the Limburgian Dialects* (WLD) and the *Dictionary for the Flemish Dialects* (WVD) inventory the vocabulary of the southern Dutch dialects. They are thematically arranged following the lexicographic idea's of A. Weijnen, which also are at the basis of still other dictionaries for some eastern dialect groups in the Netherlands. Because of their onomasiological arrangement, however, the dictionaries of Weijnen's school cannot render detailed semantic information. Therefore, professional lexicography has to call in the help of 'amateur' lexicography, i.e. the huge amount of alphabetical regional and local dialect dictionaries, made by non-professional lexicographers.
In this paper a pilot project is presented, which aims at the creation of a lexicographical database for the alphabetical amateur lexicography, including both the old alphabetical tradition of the end of the 19th / beginning of the 20th century and the new tradition, rooted in the so-called dialect renaissance of the 70s and afterwards. It is defended that such a database – if enriched with the dutchifications of the dialect headwords – will prove to be an indispensable tool for lexicological research with regard to the history of the Dutch lexicon.

> **Towards the completion of the Dictionary of the Flemish Dialects**

VAN KEYMEULEN, JACQUES AND DE TIER, VERONIQUE

3 – Reports on Lexicographical and Lexicological Projects

The Dictionary of the Flemish Dialects is a major regional dialect dictionary for the Flemish dialect area, i.e. the provinces of West and East Flanders (Belgium), Zealand Flanders (the Netherlands) and French Flanders (France). The project began in 1972 at Ghent University (Belgium). It is a thematically arranged dictionary, set up along the lines proposed by A. Weijnen for the Dictionary of the Brabantic Dialects (1960-2005) and the Dictionary of the Limburg Dialects (1960-2008), its two sister projects. It combines a dictionary with a word atlas. This paper describes the state of affairs of the Flemish Dictionary with regard to data collection, data processing, presentation and publication. (The specialised software program used for the dictionary is presented in a separate paper, in which much attention is paid to the cartographic tools).

> Österreichische Pflanzennamen. Eine Webapplikation für ein thematisches Korpus

WANDL-VOGT, EVELINE AND PIRINGER, BARBARA

3 – Reports on Lexicographical and Lexicological Projects

Im Institut für Österreichische Dialekt- und Namenlexika (DINAMLEX) befindet sich eine onomasiologisch angelegte Pflanzennamensammlung, die geschätzte 31.000 mundartliche Pflanzennamen für geschätzte 2.000 botanisch-wissenschaftliche Stichwörter enthält. Neben den wissenschaftlich-botanischen Namen wurden überregionale deutsche Standardbezeichnungen und mundartliche Pflanzennamen gesammelt. In den Jahren 2000-2005 wurden sie im System TUSTEP digitalisiert. In den Jahren 2007-2010 wurde ebd. das System dbo@ema entwickelt. Es besteht aus der eigentlichen Datenbank, die zur Speicherung heterogener Dialektdaten geeignet ist, einer öffentlichen Website, einer Desktopanwendung zur Dateneingabe und eine Javascript Applikation zur Visualisierung geografischer Daten.

Seit 2008 werden die digitalen Pflanzennamen wissenschaftlich überarbeitet. Über die Website erfolgt der Zugriff auf die Datenbank und verschiedene Contentbereiche des Systems, z.B. Lemmata, Belege, Bibliographie, Personen, Multimedia. Eine interaktive Karte, wie sie beispielsweise von Google Maps bekannt ist, stellt eine lokationsspezifische Navigationsmöglichkeit dar. Über ein Popup können die raumbezogenen Daten auch über die Karte abgefragt werden und kommt der Benutzer wieder zu unterschiedlichen Contentbereichen der Datenbank. Damit werden Fragen wie '*Welche Belege aus dem Ort XY gibt es in der DBÖ / in dbo@ema*?' oder '*Wo sagt man Gelbling zum Pfifferling*?' per Mausklick beantwortbar. Mitte des Jahres 2010 soll eine Pilotversion des Systems mit den Bezeichnungen für österreichische Pilze unter wboe.oeaw.ac.at online gestellt werden.

Durch die Einbindung wissenschaftlich-botanischer Pilznamen und die Geocodierung der Belege wird die Vernetzung mit anderen Datenbanken sichergestellt. Folgende Verknüpfungsmöglichkeiten sind projektiert und demonstrieren beispielhaft den durch Standardisierung und Geocodierung zu erreichenden Mehrwert: Verlinkung mit der Online Flora von Österreich (http://62.116.122.153/flora/Hauptseite [Access date: 14 April 2010]) und der Datenbank der Pilze Österreichs (http://austria.mykodata.net/ [Access date: 14 April 2010]).

> **The Dictionary of Lithuanian (LKŽ) and its Future in Databases and Electronic Versions**

ZABARSKAITĖ, ELENA JOLANTA AND NAKTINIENĖ, GERTRŪDA

3 – Reports on Lexicographical and Lexicological Projects

The paper deals with the Dictionary of the Lithuanian Language (Vol. 1-20, 1941–2002): electronic release, 2005 (renewed version 2008) and its new version on the CD. A three-level lexical database: exhaustive for academic purposes, medium for the broad public, and more narrow for schools, is being created at the Institute of the Lithuanian Language. Its core consists of an electronic version of the Dictionary of the Lithuanian Language (about 0.5 million dictionary entries) and its card index (about 5 million cards), which is in the process of being computerized.

> **The organization of entries in Spanish-English/English-Spanish bilingual dictionaries**

DECESARIS, JANET

4 – Bilingual Lexicography

This paper discusses the organization of equivalents and presentation of fixed expressions in six bilingual dictionaries of Spanish and English. The dictionaries studied were published over the last forty years (1971, 1983, 2003, 2004, and 2008), and we compare the information contained in the older dictionaries with more recent ones. In addition, we compare frequency data taken from the Corpus del Español with information on fixed expressions contained in the dictionary entries. The focus of the study is on the representation of the Spanish words *cuadro* and *hoyo*, and the English word *poison*. The discussion herein would be of benefit to those planning a new bilingual dictionary or major overhaul of an existing one.

> **Lexin – a report from a recycling lexicographic project in the North**

HULT, ANN-KRISTIN; MALMGREN, SVEN-GÖRAN AND SKÖLDBERG, EMMA

4 – Bilingual Lexicography

In the late 70s, the Swedish Board of Education initiated a project (the *Lexin* project) aiming at production of dictionaries between Swedish and many immigrant languages. A monolingual Swedish dictionary was compiled, serving as the common base of the bilingual dictionaries. In the 90s, the project was exported to other Nordic countries. Since the Nordic languages are closely related, much of the work carried out in Sweden could be reused in Norway, Denmark, and Iceland. Today, there are many learners' dictionaries between Nordic languages and 'exotic'

immigrant languages, especially with Swedish and Norwegian as source languages.
In this paper, we account for some aspects of this – in some respects probably unique – project. At the end, we give a description of the revision and updating of the Swedish database that has been going on since 2008

> Word-formation in English-French bilingual dictionaries: the contribution of bilingual corpora

LEFER, MARIE-AUDE

4 – Bilingual Lexicography

Research on the representation of word-formation in dictionaries is scarce and tends to be restricted to learners' dictionaries and monolingual dictionaries intended for native speakers. Nor is the issue of word-formation in bilingual dictionaries often discussed in lexicographic studies. This study, intended as a step on the way to rectifying the situation, reports the results of a comparison of the strategies adopted in four influential English-French dictionaries, focusing more particularly on derivational prefixes. The study shows that prefixes and word-initial elements in general receive very scant treatment in English-French dictionaries, which seems hardly justifiable when one thinks of the major role they play in the interpretation and translation of complex words. In my presentation I will highlight and illustrate a number of shortcomings, such as the lack of consistent criteria for the selection of affix entries and the misrepresentation of affix polysemy. More importantly, the presentation will also show how bilingual dictionary-making could benefit from bilingual corpora (both comparable and translation corpora) to improve the description of word-formation. I will propose a corpus-based list of the most productive and frequent prefixes in English and French. This list would seem to be a promising starting point for selecting more systematically and more rigorously the affixes to be included as headwords in bilingual dictionaries. To illustrate the usefulness of corpus data, I will also present a model bilingual entry for the French prefix *dé–* based chiefly on data extracted from an English-French translation corpus.

> Problems of Dialect Non-Inclusion in Tshivenḓa Bilingual Dictionary Entries

MAFELA, MUNZHEDZI JAMES

4 – Bilingual Lexicography

Language is human speech involving the use of words in an agreed way. However, a language is not absolutely homogeneous since there is

variation. In any language one can expect to come across instances where certain speech differences may exist due to the influence of a language in an adjoining area. Tshivenḓa is characterised by a number of dialects, among them Tshiphani, Tshiilafuri, Tshimbedzi, Tshironga, Tshimaanḓa and Tshinia, which exhibit some linguistic features different from those of other groups. The standard dialect in Tshivenḓa is Tshiphani. This dialect is spoken in the areas of Tshivhase and Mphaphuli. The selection of the Tshiphani as a standard dialect in Tshivenḓa did not cause the other dialects to die out as they are still used by the Vhavenḓa as spoken language. However, there is non-inclusion of dialectal entries in some dictionaries, whereas in others, very few dialectal entries have been included. Some dialects differ from the standard dialect in vocabulary, whereas others differ from the standard dialect in pronunciation.
A lexicographer must always take into consideration that there is a variation in language. Lexicographers should not see the inclusion of non-standard dialects in a dictionary as corrupting the standard language. The inclusion of non-standard dialects in dictionaries, especially bilingual dictionaries, will assist dictionary users to know more about variants in the language. A dictionary is expected to accommodate all dialects of a language because they have equal value in spoken language. It is important for a lexicographer to first carry out research regarding the existence of dialects in a language if one intends to compile a dictionary. This paper seeks to show that it is necessary to include lexicons from non-standard dialects in lexicography works such as bilingual dictionaries because there is no dialect which is better than others. The addition of non-standard dialects in dictionaries will enrich the languages.

> **Approche historique et sociolinguistique de la lexicographie bilingue missionnaire et les langues minoritaires en Algérie coloniale (1830-1930): le cas du berbère**

MAHFOUD, MAHTOUT AND GAUDIN, FRANÇOIS

4 – Bilingual Lexicography

Notre propos prend place dans le cadre de l'histoire culturelle des dictionnaires. Nous nous proposons de mettre en lumière les circonstances qui expliquent et déterminent le développement de la lexicographie bilingue missionnaire dans l'Algérie colonisée. Nous traiterons plus particulièrement du cas du berbère.
La création, en 1868, de la Société des missionnaires d'Afrique en Algérie marque une nouvelle étape dans l'action missionnaire africaine. Depuis leur installation, les missionnaires ont œuvré pour faire sortir de l'anonymat la langue minoritaire du peuple berbère.

Le point de départ de notre étude relève d'un constat: même si les instructions, claires et rigoureuses, des supérieurs de la mission exigeant de leurs missionnaires une étude assidue et une connaissance approfondie de l'arabe, nous constatons que leur production lexicographique n'inclut aucun dictionnaire bilingue ayant pour objet la langue arabe. Or, toute la production lexicographique des missionnaires porte sur les différents dialectes berbères. Dès lors, cette orientation de la lexicographie-missionnaire ne manque pas de soulever des questionnements : a) Quelles étaient les instructions données aux missionnaires concernant l'étude des langues locales? b) Pourquoi les missionnaires se sont-ils penchés sur les différentes variétés berbères plutôt que sur l'arabe algérien alors que c'était la langue d'intercompréhension entre les communautés indigènes? La lexicographie bilingue-missionnaire a-t-elle participé à la valorisation des langues minoritaires berbères?

Le zèle missionnaire s'est centré sur la langue des berbères. Soumis aux choix de leur hiérarchie, influencés par le mythe berbère que cultive Lavigerie, confrontés aux nécessités de leur travail d'évangélisation, les missionnaires ont utilisé la langue berbère comme instrument pour répandre la bonne parole.

En composant des outils fondamentaux pour la vulgarisation de la langue berbère, les missionnaires ont contribué à la grammatisation de la langue berbère et au recueil d'un lexique devenu précieux pour les études sociolinguistiques.

> **The TRANSVERB project – An electronic bilingual dictionary for translators: theoretical background and practical perspectives**

SÁNCHEZ CÁRDENAS, BEATRIZ AND TODIRASCU, AMALIA

4 – Bilingual Lexicography

TRANSVERB is a lexicographic resource conceived for novice and professional translators who need assistance when translating texts into a foreign language. It is a semi-bilingual dictionary which can also be used for text production into a foreign language. The case study analyzed in this article pertains to the translation of verbs from French to Spanish. This dictionary is organized onomasiologically in terms of categories. Based on the hypothesis that human cognition organizes concepts in semantic categories (Tranel et al. 2001; Damasio et al. 2004), TRANSVERB is configured in lexical domains (Martín Mingorance 1985, 1987, 1900, 1995; Faber & Mairal 1999). The syntactic information in verb entries includes its combinatory potential, more specifically, its number of arguments as well as their semantic restrictions. This is established through corpus study.

> **The retrieval of data for Slovene-X dictionaries**

ŠORLI, MOJCA

4 – Bilingual Lexicography

The article reflects on the linguistic issues concerning the preparation of text for a new Slovene-English dictionary. Discussion is based on concrete examples from the reversed database of the Oxford-DZS Comprehensive English-Slovene Dictionary (2005/2006) and lexicogrammatical data from a corpus-based Slovene lexical database in the making. It is yet to be established how successful retrieval of data from a reversed bilingual database is. However, the first attempts to use information from the reversed database for the purposes of the compilation of a new Slovene-English dictionary indicate that the automatically generated database is a vast fund of information on the contrastive relations between English and Slovene, which should at no cost be overlooked. The user has instant access to the potential direct translation candidates, and to the more contextually-bound potential translations, many of which would have been inaccessible to a non-native speaker without an insight into what could be called the mirror image of the language. On saying that, it is important to stress that a reversed database as we understand it is in no way to be confounded with the actual 'reversed' dictionary itself, but merely to be seen as a bilingual framework in which no solution is automatically transferred to a Slovene-English dictionary. We come to the conclusion that while a corpus-based monolingual database is needed to provide a fresh and authentic image of the source language, it is also important to explore and exploit the data obtained in the reversed bilingual database because that will add an extra dimension to the Slovene-X dictionary text. The key question is how to proceed with the compilation of a new, bilingual, dictionary database, using both sources but avoiding a distorted lexical analysis of Slovene in use, while also ensuring a thorough contrastive analysis of the relationships between the two languages.

> **OMBI bilingual lexical resources: Arabic-Dutch / Dutch-Arabic**

TIBERIUS, CAROLE; AALSTEIN, ANNA AND HOOGLAND, JAN

4 – Bilingual Lexicography

In this paper we present the OMBI reversible bilingual lexical resources for Dutch-Arabic and Arabic-Dutch. These resources have been derived from a bilingual lexical database which has originally been produced with OMBI, a special tool for creating and editing bilingual dictionaries, within the framework of the project 'Woordenboek Nederlands-Arabisch,

Arabisch-Nederlands, Nijmegen' in the period of 1998 till 2002 at the Radboud University of Nijmegen. Printed dictionaries have been published on the basis of this database (Hoogland et al. 2003) and now the data has been converted to LMF (Maks et al. 2008) to ensure future interchangeability and interoperability.
OMBI-Arabic-Dutch and OMBI-Dutch-Arabic are part of a larger set of bilingual lexical resources which are available at the Dutch HLT Agency. The main strength of these bilingual computational resources is the high quality of the input data, which exceeds that of most existing computational resources, since it is based on the work of a team of professional lexicographers. In addition, most of these bilingual resources use the same Dutch component as a base, which offers interesting perspectives for linking the resources to each other following the hub and spoke model.

> **Reversing a Bilingual Dictionary: a mixed blessing?**

VELDI, ENN

4 – Bilingual Lexicography

The presentation focuses on the experience of reversing a general Estonian-English dictionary of about 49,000 entries and 93,000 equivalents by means of the Tshwanelex dictionary compilation software. The reversal served two purposes. First, it seemed appropriate to reuse the established cross-linguistic equivalents in the Estonian-English dictionary for the B part of a new English-Estonian dictionary. Second, one also expected to enlarge and improve the reversed Estonian-English dictionary in the course of the post-editing phase. So far the post-editing phase of the English-Estonian dictionary has been highly rewarding. In fact, it could be regarded as simultaneous cross-fertilization of both dictionaries, especially with regard to additional meanings and a more balanced treatment of synonyms. On the other hand, the post-editing phase of a general dictionary has been more time-consuming than expected. It is also argued that, on the one hand, the reversal mercilessly reveals the drawbacks of the B part of a bilingual dictionary, such as explanation-like equivalents, inaccurate equivalents, lexical poverty, etc. In fact, it appears that many dictionaries are not actually suitable for reversal. On the other hand, in the case of reversibly oriented dictionaries the post-reversal editing process may result in enriched target and source dictionaries – and will considerably reduce asymmetry in bilingual dictionaries.

> **Management and use of terminological resources for distributed users in the translation hosting site Minna no Hon'yaku**

ABEKAWA, TAKESHI; UTIYAMA, MASAO; SUMITA, EIICHIRO AND KAGEURA, KYO

5 – Lexicography for Specialised Languages – Terminology and Terminography

In this demonstration, we show the terminology management module of Minna no Hon'yaku (MNH: http://trans-aid.jp/), a translation hosting site with integrated translation-aid mechanisms, which was made publicly available in April 2009. As of February 8th, 2010, 1062 users have registered with MNH and more than 3400 documents have been translated, of which more than 1600 translations have been published on the site. On MNH, users can translate documents individually or can define groups and share the translation task. It provides users with functions such as lookup of high-quality dictionaries and terminologies, seamless access to Wikipedia and Google search, and reference to TM. There are two types of terminological resources on MNH, i.e. those provided by the system and those registered by users. The demonstration shows how terms are registered, shared and used.

> **Adjectives and collocations in specialized texts: lexicographical implications**

ALONSO CAMPOS, ARACELI AND TORNER CASTELLS, SERGI

5 – Lexicography for Specialised Languages – Terminology and Terminography

The *General Theory of Terminology* (Wüster 1979) states that all terms must be nouns, as the noun is the only category to designate a concept. For this main reason, adjectives and other grammatical categories are not considered as entries in most terminological dictionaries. The *Communicative Theory of Terminology* (Cabré 1999, 2000, 2002), on the other hand, has recently determined that predicative categories, such as adjectives, verbs and adverbs, can also become specialized lexical units (SLU). However, there are not enough empirical studies at this moment which confirm this hypothesis and examine the main characteristics of these predicative categories when they are used as terms. Specifically, our contribution studies the use of adjectives as terminological units in environmental texts. The study of Environment-related terminology is of special interest, as Environment is a new emerging domain with characteristics different from those of classical domains, such as Medicine or Chemistry. As it has been established in previous works (Alonso 2009, Bracho 2004), many Environment-related words are taken from the general language, but take on a terminological sense when they are used in environmental texts.

Our study focuses on adjectives which form a collocation $[N[A]_{SAdj}]_{SN}$, as this syntactic structure is frequently used in specialized discourse. Our main objective is to determine the 'terminological value' of these adjectives and their main characteristics. It is concluded from the data analysis results that the behaviour of adjectives depends mainly on the syntactic-semantic nature of the adjective. It is observed a general tendency to use as terms, either classifying relational adjectives (Bosque 1993, Bosque & Picallo 1996, Picallo 2002), or common qualifying adjectives that adopt a terminological sense in specialized texts. This fact brings about the need of different kinds of treatment for the representation of these adjectives in terminological dictionaries.

> **'Not Leaving Your Language Alone': Terminology Planning in Multilingual South Africa**

BEUKES, ANNE-MARIE

5 – Lexicography for Specialised Languages – Terminology and Terminography

Status language planning has been one of the components of post-apartheid South Africa's transformation project that has managed to attract wide-spread attention. In 1994 South Africa moved from its former official bilingual language policy to a new constitution that enshrines official status to 11 of the languages spoken in South Africa. However, 16 years down the line there is widespread disappointment with organized language planning and management by government authorized agencies. The paper gives a brief analysis of terminology development in contemporary South Africa juxtaposed with a terminology development project at the micro level which, in Joshua Fishman's words, was initiated from the perspective of 'not leaving your language alone'.

The practice of translation is an age-old activity, but translation studies is a fairly 'new' academic discipline and hence its terminology is still in its infancy. Translation studies has been taught in South Africa at higher education institutions for more than thirty years, but mainly through the medium of English and Afrikaans. The prod for this project was therefore the identification of fresh needs for terminology development in this area to contribute to facilitating the sustained development of specialized discourses in higher education. Terminology development is viewed as indispensable for creating and sustaining a dynamic environment for the use of South Africa's official indigenous languages as a medium of instruction and ultimately for scientific progress.

> **Extension of a Specialised Lexicon Using Specific Terminological Data**

CARTONI, BRUNO AND ZWEIGENBAUM, PIERRE

5 – Lexicography for Specialised Languages – Terminology and Terminography

The paper describes methods for acquiring lexical information to implement a 'Unified Medical Lexicon for French' (UMLF) that aims at being a reference resource for NLP in the medical domain. We address four issues of lexical acquisition in a specialised domain. First, to assess the 'desired coverage' of lexical information, we use a large collection of French terms as a reference resource for the medical domain sublanguage. The collection contains close to 300,000 terms organised around conceptual identifiers. Second, by looking through this large amount of terminological data, we highlight the different kinds of information that might be useful to deal with typical terminological processing tasks, like variant recognition. The terminological variation phenomena that are very frequent in these terms are of three kinds: graphemic, inflectional and derivational variations. Third, we propose a model for organising the lexical information. Most of this model is inspired from existing specialist lexicons, but special emphasis is put on derivational morphological information. Finally, different kinds of acquisition methods are described, at the two levels of linguistic description that are addressed here: inflectional and derivational morphological knowledge. These methods allow acquiring an important amount of lexical data. For inflectional knowledge, the full paradigm is recorded, to provide information about all the possible inflected forms of lexical units within terms. Regarding derivational knowledge, specific derivation processes are targeted, in order to handle particular term variations. The relevance of the gathered derivational information is also assessed.

> **Bilingual Technical-Translation Thesaurus as a Reliable Aid to Technical Communication**

FAAL HAMEDANCHI, MARYAM

5 – Lexicography for Specialised Languages – Terminology and Terminography

The article reviews the problem of technical terminology translation and the role it plays in technical communication. Despite the progressing attempts for standardization of terminology, there is still long distance to a perfect terminology system practically in all language societies. For an individual concept are used different variants even within a single text and technical dictionaries often fail to cover all these variants. Languages do not possess the same instruments for illustrating a definite concept, as a result, in translating different equivalents of a single concept the translated terms

may be considered as synonyms rather than variants or, on the contrary, partial synonyms of a term in the source language can be considered as variants or close synonyms in the target one. The problem gets even more complicated when it comes to languages, namely Persian and Russian, where the users are imposed to employ English as an intermediate language.
Technical dictionaries pay less attention to these differences, at the best, they may provide scope notes or short definitions to distinguish different senses of a term, which hardly suffices for a proper communication. On the other hand, users of a bilingual technical dictionary may look up different kinds of information besides definition and equivalents. They may look up cross-language synonymous or antonymous, allocations, homonyms and other information, which are rarely provided by a bilingual technical dictionary.
These facts imply the necessity of employing more onomasiological approach in compiling bilingual technical dictionaries. In our opinion, a revised structure of information-retrieval thesauri complies in a better way with the requirements of technical dictionaries.
A technical-translation thesaurus can reveal the basic structure of an information retrieval thesaurus, but compiles the necessary features of a common language thesaurus and provide approaches to equivalents of a term in different languages starting from the concept, which does not depend on the language.

> **Introducing the Dutch Terminology Service Centre: a centre of expertise on practical terminology work**

GÖRÖG, ATILLA

5 – Lexicography for Specialised Languages – Terminology and Terminography

The Dutch Terminology Service Centre (DTSC, Steunpunt Nederlandstalige Terminologie) was founded in 2007 by the Dutch Language Union, a Dutch-Flemish government institution. The DTSC functions as a non-commercial information center for all aspects of terminology and serves the entire Dutch-speaking community. We give advice on terminological research to anyone who is involved in terminology-related work (companies, organizations, translators, terminologists, teachers, scientists etc).

> **The Tension between Definition and Reality in Terminology**

HACKEN, PIUS TEN

5 – Lexicography for Specialised Languages – Terminology and Terminography

Whereas natural language concepts are based on prototypes, classical terminological definitions (CTDs) are based on necessary and sufficient conditions. The sociocognitive approach to terminology rejects both the

possibility and the desirability of CTDs. In this paper, I argue that CTDs are needed for certain types of term, but not for others.
In a first step, I distinguish two overlapping classes of expressions, based on two different criteria for termhood. Specialized vocabulary is the set of items whose use is restricted to specialized communication. Specialization is a gradual property, so that the boundaries of this class are vague. TERMS in the narrow sense have a concept with a clearcut boundary. In scientific and legal contexts, clearcut boundaries are necessary because classification is important. Only for TERMS in this narrow sense do we need CTDs.
Some of the problems raised for CTDs can be solved straightforwardly by restricting their domain to TERMS. Discussions about the best definition, e.g. for governing category in Chomskyan linguistics, indicate a search for the best concept rather than a prototype nature. In cases such as significant in statistics, the CTD is logically independent of the prototype concept associated with the word significant in general language. In such cases, finding CTDs does not pose insuperable problems.
More difficult problems arise when scientific concepts interact with our intuition about classification and with technological advances. The discussion of species in biology, compound in linguistics, and planet in astronomy demonstrates how these problems arise and how they can be addressed without recourse to prototype-based concepts. TERMS in the narrow sense may be unnatural and their definition not straightforward, but they are crucial in scientific communication and depend on CTDs.

> **Termania – Free On-Line Dictionary Portal**

KREK, SIMON

5 – Lexicography for Specialised Languages – Terminology and Terminography

Termania, a free on-line dictionary portal with integrated dictionary browsing and editing tools is being developed by Amebis software company from Kamnik, Slovenia, in cooperation with Trojina, Institute for Applied Slovene Studies. It provides an interface for dictionary browsing and a simple but reasonably versatile on-line dictionary editing tool. The portal is intended for general public users with no specialized computer or lexicographic knowledge, but with an interest to share terminological or general language knowledge, either by offering translations in a bilingual or multilingual environment or providing definitions in a monolingual context.
The portal is intended to serve as the central terminology data and opinion exchange node for Slovene terminology. Therefore, a discussion forum will be included in the portal and a possibility to comment on and evaluate each particular entry. Internationalization of the portal is

foreseen by providing language-specific interface for all widely used languages. General policy for the use of dictionary data on and off the portal is determined by the Creative Commons Attribution Non-Commercial Share Alike licence but also other licences can be used, according to users' preference.
The dictionary portal will be available at the web address: http://www.termania.net/ from July 2010.

> TermFactory: A Platform for Collaborative Ontology-based Terminology Work

KUDASHEV, IGOR; KUDASHEVA, IRINA AND CARLSON, LAURI

5 – Lexicography for Specialised Languages – Terminology and Terminography

TermFactory is an array of standards and tools based on Semantic Web ontology techniques. Its mission is to allow companies, organizations and individual contributors to collaboratively produce multi-domain special language vocabularies and ontologies.
Ontologization of terminological data has several benefits, such as global identification of concepts, automatic checks for logical errors, reasoning and data propagation, presentation of data in machine readable and -processable form and the possibility to substitute static entries with dynamic 'views' tailored according to the user's needs and preferences.
Collaborative work is a double-edged sword which potentially has many benefits but may also present serious challenges. In our poster, we describe challenges of collaborative terminology work and possible solutions to them. If well-organized, a collaborative project can be quite successful, as the example of Wikipedia and many other collaborative projects on the Internet demonstrate.

> The Focal.ie National Terminology Database for Irish: software demonstration

MĚCHURA, MICHAL BOLESLAV AND Ó RAGHALLAIGH, BRIAN

5 – Lexicography for Specialised Languages – Terminology and Terminography

In this demonstration, we will showcase the National Terminology Database for Irish focal.ie which was launched on-line in 2006 and immediately became very popular, attracting hundreds of thousands of searches every month. The Web site allows users to search a database of around 300,000 terms. It was developed to make the stock of terms of the National Terminology Committee available online, and was designed to be easy to use.
In addition to the public-facing Web site, the software comprises an

editorial interface (password-protected Web site), a relational database, and a library of objects and functions that acts as an interface between the two Web sites and the database.
The public-facing Web site allows users to search the database using a 'Quick Search', a 'Complex Search', or an 'Alphabetical Listings' function. The 'Quick Search' function returns 'Similar terms', 'Exact matches', and 'Related matches' from the database. The editorial interface allows users to search and edit the data contained in the database.
While the primary requirement for the public-facing Web site is user-friendliness, the primary requirement for the database is the ability to record complex linguistic data in a logical structure. The structure adopted for the Irish lexical database is based on the conceptual model, widely considered a standard in the terminology industry (ISO 704). The database is multilingual and contains rich grammatical labelling, usage examples, definitions, as well as other information. The database, editorial tools, and public interface were developed by Fiontar in-house using Microsoft technologies. The database and Web sites are hosted by Information Systems & Services, Dublin City University.
A new version of the public site was recently launched and can be accessed at the following URL: http://www.focal.ie/

> **Towards a bilingual lexicon of information technology multiword units**

MOSZCZYŃSKI, RADOSŁAW

5 – Lexicography for Specialised Languages – Terminology and Terminography

The article presents a proposal of an electronic, English-Polish translation dictionary covering the language of computer science. The dictionary will focus on multiword units and phraseology typical for this domain. It is supposed to answer the needs of technical translators, who can easily access simple terminological databases, but lack good production dictionaries that would go beyond single terms. The proposed dictionary aims at filling this gap by focusing on multiword units and their modifications, as well as on individual terms' collocational patterns.
The dictionary will be based on the idea of 'extended phraseology' proposed by Müldner-Nieckowski. According to this idea, phraseology is not limited to idioms in the traditional sense of the word, but also covers phrasemes (i.e. units with conventionalized structure, but without figurative meaning), as well as phraseograms (syntactically incomplete units that carry some semantic value). Such a broad approach to phraseology in the planned dictionary will allow translators to create texts that sound natural to computer science experts and to maintain consistency on the stylistic level on top of terminological consistency.

The dictionary will be created in electronic form, with the aim to make it available free of charge on the Internet as part of the Freedict project.

> **Building on a terminology resource – the Irish experience**

NIC PHÁIDÍN, CAOILFHIONN; Ó CLEIRCÍN, GEARÓID AND BHREATHNACH, ÚNA

5 – Lexicography for Specialised Languages – Terminology and Terminography

www.focal.ie is the national database of Irish language terminology. In this paper, we examine: (i) the impact achieved by this resource in the five year period since work commenced; (ii) the possibilities which have arisen from one project over a short time span, to develop sub-projects and related initiatives; and (iii) the advantages and opportunities arising from the creation of one high-quality electronic language resource. The Irish case shows that the development of high-quality resources for a lesser-used language can have interesting and unexpected knock-on effects.

We present eight stages and aspects of term planning: preparation/planning; research; standardisation; dissemination; implantation; evaluation; modernisation/maintenance; and training. Fiontar, in its work, has moved from its initial involvement in the *dissemination* of terminology, to take an active part in other aspects of term planning for Irish: *research, standardisation, evaluation, modernisation* and *training*. This has been achieved through *editorial and technological development*, in *partnership* with key stakeholders and always from a *socioterminological* point of view – that is, with an emphasis on terminology as an aspect of language planning and from the point of view of users in particular.

Particular projects described include Focal as a term management system and as a user resource; tools for translators; user links to a corpus; the development of a new sports dictionary; and research into subject field headings. Two related projects are the LEX legal terms project for term extraction and standardisation, and the development of terminology for the European Union.

> **L'ancien et le moyen français au siècle classique: le *Tresor de Recherches et Antiquitez Gauloises et Françoises* de Pierre Borel (1655)**

AMATUZZI, ANTONELLA

6 – Historical and Scholarly Lexicography and Etymology

Pierre Borel est 'le premier des savants qui se mirent à rédiger des recueils où les mots de l'ancienne langue étaient consignés' (G. MATORÉ, *Histoire des dictionnaires français*, Paris, Larousse, 1968, p. 132). En effet son

Tresor de Recherches et Antiquitez Gauloises et Françoises est un ouvrage qu'on peut compter parmi les premiers dictionnaires d' 'ancien français' et qui doit être également signalé pour les préoccupations étymologiques qu'il affiche et parce qu'il recense beaucoup de régionalismes.
Dans le cadre de ce colloque je focalise mon attention sur la manière dans laquelle Borel affronte la problématique de l'évolution du français et en particulier sur la façon dont ce dictionnaire présente et envisage la langue ancienne.
Dans sa préface longue et articulée, qui contient un discours métalinguistique explicite, Borel nous livre des considérations à propos du processus de transformation qu'investit les langues. Il s'agit des réflexions d'un homme cultivé du Grand Siècle qui entend valoriser la 'langue ancienne' et qui insiste sur le fait que l'évolution linguistique est tout à fait naturelle. Il affirme une vision de la langue où diachronie et synchronie sont complémentaires et indissociables, ce qui se traduit dans son souci constant d'aborder les phénomènes linguistiques (le lexique mais aussi la structure de la phrase et l'évolution phonétique) dans une perspective historique et culturelle.
À une époque où le français subit une épuration sévère et les archaïsmes sont proscrits, Borel va contre courant. Il produit un dictionnaire de la langue et de la civilisation anciennes,. utile pour accéder à la culture gauloise et française du Moyen Âge et pour comprendre ses évolutions successives, qui constitue un témoignage précieux de la valeur culturelle des mots, de la richesse et de la vivacité d'expression tant de l'ancien que du moyen français. Mais le *Tresor*, qui atteste aussi du développement du vocabulaire jusqu'au français classique, devient également 'mémoire' de comment la langue et la civilisation se sont construites le long des siècles.

> La compilation de dictionnaires de synonymes distinctifs: une démarche synonymique et lexicographique

FERRARA, ALICE

6 – Historical and Scholarly Lexicography and Etymology

En 1718 naissait un nouveau genre de dictionnaire: le dictionnaire de synonymes monolingue français appelé *a posteriori* dictionnaire de synonymes distinctif (par opposition aux dictionnaires de synonymes cumulatifs que nous connaissons aujourd'hui.)
Les dictionnaires distinctifs se composent de définitions, précisions, exemples qui justifient les propos de l'auteur. Quand les dictionnaires de synonymes se sont multipliés, à peine un siècle après la naissance du genre, apparaît un genre annexe, celui de compilation. Le but des auteurs de compilations était de garder ce qu'ils jugeaient être le

meilleur des dictionnaires de synonymes les précédant. Dans notre article, nous nous interrogerons sur la place des compilateurs et si l'on peut dire que ce sont eux-aussi des synonymistes. Nous montrerons que compiler les dictionnaires de synonymes c'est faire preuve d'autant de travail synonymique et lexicographique que de composer directement le dictionnaire, car pour compiler des dictionnaires de synonymes il faut faire des choix multiples. Tout d'abord il faut choisir les auteurs dont on souhaite compiler les dictionnaires. Ensuite il faut sélectionner les articles à reprendre, ce qui représente un véritable choix du compilateur par rapport aux termes qu'il estime être réellement synonymes. Et enfin, il faut étudier ce qu'un compilateur garde des articles préalablement choisis. En effet, il gardera d'un article ce avec quoi il est en accord, et inversement, retirera tout ce à quoi il n'adhère pas. De plus, outre tous ces choix, le compilateur se fait également entendre puisqu'il peut aussi composer lui-même un certain nombre d'articles. Nous pouvons donc dire que la compilation de dictionnaire de synonymes distinctif est un véritable travail de réécriture fondé sur l'écriture et la création puisque les compilateurs n'ont de cesse de faire des choix lexicographiques.

> **Lexicography, Printing Technology, and the Spread of Renaissance Culture**

HANKS, PATRICK

6 – Historical and Scholarly Lexicography and Etymology

Historians of lexicography in the English-speaking world have implied that Robert Cawdrey's *Table Alphabeticall* (1604) is the first English dictionary. Landau (1984, 2001) makes this claim, adding that it is 'the least inspiring of all seminal works'. In this paper, I agree that the *Table Alphabeticall* is uninspiring, but I deny that it is a seminal work. Landau overlooks the rich 16th-century tradition of Renaissance and Humanist lexicography in Europe, in particular the *Thesaurus Linguae Latinae* of Robert Estienne (1536) and the *Thesaurus Linguae Graecae* of his son Henri Estienne (1572). These seminal works are astonishing achievements—breathtaking innovations—in terms of both scholarship and technology. They set standards for subsequent European lexicography. Two technological innovations made these great dictionaries possible: the invention of printing by Gutenberg in Strasbourg in about 1440 and the typography of Nicolas Jenson in Venice in 1462. These technological developments and the lexicographical achievements that were made possible by them contributed, in the first place, to the Renaissance programme of preserving the classical heritage of ancient Greece and Rome and, in the second place, to the role of dictionaries in spreading Renaissance culture and Humanism across Europe. The paper goes on to briefly outline the emergence of bilingual lexicography,

replacing the polyglot lexicography that was standard in the 16th century. A comparison is made between the influence of printing technology on 16th century lexicography and the potential influence of computer technology on 21st century lexicography.

> Grammatical information in dictionaries

HOEKSTRA, ERIC

6 – Historical and Scholarly Lexicography and Etymology

A dictionary is an encyclopaedia of linguistic information about words. It presents to a target group of laymen and professionals general information about words belonging to various disciplines of linguistics such as

- semantics (the meaning of words and phrases)
- phonology (the pronunciation of words)
- syntax (the syntactic category of words and the collocations in which they partake)

My contribution discusses the question whether (new) insights from these disciplines may change the content of dictionaries, seeing that an evaluation of these insights does not take place very often.

It is a shortcoming of dictionaries that a paraphrase of the meaning of function words is often not very insightful with respect to their use (Coffey 2006).

- What is the meaning of Dutch *er* 'there'?
- What is the meaning of articles like *the*?
- What is the meaning of the complementiser *that*?

Some Dutch dictionaries muddle the description of the various uses of *er*, ignoring the distinctions drawn by de *Algemene Nederlandse Spraakkuns*t, the Standard Dutch Grammar (Haeseryn et al. 1997). Those distinctions are practical and well-motivated (Hoekstra 2000). It is proposed to use syntactic knowledge to structure articles about function words. In addition, dictionaries can covertly use example sentences to illustrate syntactic phenomena. Such measures strengthen the encyclopaedic character of a dictionary.

> Celtic Words in English Dictionaries and Corpora

ITO, MITSUHIKO

6 – Historical and Scholarly Lexicography and Etymology

The researcher has collected Celtic words from several English dictionaries and a few English etymological dictionaries and found that there are about 300 words in present English dictionaries. He has studied what words and how many of the 300 words native speakers of English know. The

research method was giving matching tests of words and definitions and having subjects write appropriate words to definitions. The subjects were all adult voluntaries. Main purposes of the present study are: (1) to survey what words of the 300 words appear in BNC and Wordbnaks, and (2) to survey if well known words by native speakers are highly frequent words in BNC and Wordbanks. Two main results have deduced from the present study. One is that not all of the 300 words appear in BNC and Wordbanks and some words appear in the two Corpora and some others appear in either of the Corpora and the others do not appear in both of them. The other is that well known words in the research do not necessarily come to the top frequent positions of the Corpora.

> **Annotations in *Dictionarium Latino Lusitanicum, ac Iaponicum* (1595) in the Context of Latin Education by the Jesuits in Japan**

KISHIMOTO, EMI

6 – Historical and Scholarly Lexicography and Etymology

The Jesuits in Japan began establishing schools in the 1580s to mentor young native men in priesthood. In 1594, their students received a printed abridged edition of the Latin grammar, originally written by Manuel Alvarez, and the next year they received *Dictionarium Latino Lusitanicum, ac Iaponicum* (DLLI), a Latin-Portuguese-Japanese dictionary based on the Latin dictionary compiled by Ambrogio Calepino.

One of the features, when comparing the DLLI with the original, is that it cites the names of Latin classical writers without quoting sentences in several entries. This paper attempts to clarify the reasons for these annotations in this edition and reflects on the purpose of the DLLI.

Plautus is cited in about 70 entries, the most citations among all the names found in the DLLI. However, this number does not reflect the number in the original, which includes many classical writers, especially Cicero, whose works were regarded as a model for Latin prose. We also have no evidence showing that Jesuits in Japan regarded Plautus's writing as more important than Cicero's in teaching Latin.

The editors of the DLLI cite Vergilius most frequently after Plautus; we also find many annotations from the original showing the differences in usages such as 'apud veteres' (used by ancient people) or 'apud poetas' (used by poets). Similarly, it is reasonable to suppose that the editors included notes on 'Plaut' to describe the differences in older usages. They appear to retain the citations of writers and other annotations on special usages in order to teach the various nuances of Latin vocabulary to students in Japan, many of whom had elementary or intermediate language skills and needed good Latin proficiency to work as priests.

> **Vers un enrichissement raisonné de la rétroconversion du *Französisches Etymologisches Wörterbuch* (FEW)**

MAZZIOTTA, NICOLAS AND RENDERS, PASCALE

6 – Historical and Scholarly Lexicography and Etymology

L'informatisation par rétroconversion du Französisches Etymologisches Wörterbuch (FEW) de Walther von Wartburg, oeuvre fondamentale de la lexicologie historique galloromane, est à présent en cours de réalisation et nous voudrions montrer en quoi elle est perfectible et comment elle pourrait être améliorée. Nous présentons brièvement le FEW et la nécessité de le rendre exploitable par l'ordinateur (attentes des utilisateurs et besoins de formalisation), puis nous exposons les principes qui ont gouverné sa rétroconversion (au format XML) et donnons l'exemple détaillé de l'article substantivus, dont nous faisons la lecture suivie. Ensuite, nous focalisons l'exposé sur la microstructure rétroconvertie des articles et l'enrichissement par annotation manuelle que nous préconisons. Nous montrons quelles informations ne sont pas accessibles à la machine en raison de l'omniprésence de conventions implicites dans l'oeuvre originale. Nous synthétiserons enfin le potentiel de notre approche: l'accès à l'intelligence du FEW, et non plus seulement à sa forme.

> **The negation particle *ne* in the historical dictionaries of Dutch**

MOOIJAART, MARIJKE

6 – Historical and Scholarly Lexicography and Etymology

In the historical stages of the West Germanic languages *ne* has been part of the negation system. In Old Dutch through Early New Dutch (c. 600 – 1600/1700) this particle had various specific functions and senses, depending on the sentence structure and on whether or not it co-occurred with other negation elements, such as negative indefinites and adverbs. The present paper focuses on the way in which this particle as function word is described in the four successive historical dictionaries of Dutch. As these dictionaries were compiled in different periods and on different editorial principles, one can expect differences in treatment of the particle. Sometimes shortage of material plays a role. The focus on translation in Modern Dutch rather than on a precise grammatical analysis, causes inadequate descriptions in other cases. Especially with respect to the conjunctional construction with *ne*, the lexicographer is in need of clear, insightful discussions of this complicated phenomenon in the linguistic literature as a basis to his description. In spite of these shortcomings, the dictionaries together contain a comprehensive survey

of the various uses of *ne*, in some of them with a detailed inventory of contexts, together with a large amount of mostly dated illustrative citations.

> **Antedating headwords in the third edition of the OED: Findings and problems**

PODHAJECKA, MIROSŁAWA

6 – Historical and Scholarly Lexicography and Etymology

The present paper describes problems involved in antedating headwords in the third edition of the Oxford English Dictionary (OED3). One of the elements in urgent need of revision in the previous edition was the dating of quotations but, despite the intensive labours of the lexicographers, some OED3 headwords and senses are still likely to be misdated. This paper is based on the premise that Google Books, a gigantic online resource, can be applied successfully for the verification of OED3 dating. Indeed, my findings indicate clearly that Google Books has a vast research potential, because half of the words in my sample (covering 129 items related to dancing) have been antedated in full-text sources. However, neither the search procedure nor the interpretation of data retrieved is straightforward, so a number of ambiguous and problematic examples have been provided to show that antedating is far more intricate than it seems at first sight. For example, one has to repeatedly distinguish between related senses, evaluate the relevance of similar word-forms and determine whether or not the words can be treated as fully-fledged loanwords. Even though many of my decisions were purely intuitive, I nonetheless hope that at least some of the antedatings found in Google Books turn out to be helpful for OED3 lexicographers in the on-going revision process.

> **Le grand vocabulaire François, un ouvrage taxé de tous les maux**

REY, CHRISTOPHE

6 – Historical and Scholarly Lexicography and Etymology

Notre communication propose une présentation du *Grand Vocabulaire François* (1767-1774), la première entreprise lexicographique du grand éditeur Charles-Joseph Panckoucke.

Nous esquissons ici les traits d'un ouvrage qui en dépit de son originalité scientifique et de ses choix linguistiques très intéressants est resté dans l'ombre de l'*Encyclopédie* et du *Dictionnaire Universel* de Trévoux.

Our paper provides an overview of the *Grand Vocabulaire François* (1767-1774), the first lexicographical work of the great publisher Charles-Joseph Panckoucke.

We present here a dictionary that despite its scientific originality and its very interesting linguistics choices remained in the shadow of the *Encyclopédie* and the *Dictionnaire Universel* of Trevoux.

> **Frauen**
> **Rollentypen in einem dialektlexikographischen Jahrhundertprojekt (1911-2010)**

WANDL-VOGT, EVELINE

6 – Historical and Scholarly Lexicography and Etymology

Das Wörterbuch der bairischen Mundarten in Österreich (WBÖ) ist ein lexikographisches Großprojekt (Sammlung seit 1911-; Publikation seit 1963-; Digitalisierung der Datenbestände seit 1993-).
In der Geschichte des Wörterbuchunternehmens lässt sich feststellen, dass Frauen eine besondere Rolle eingenommen haben, die in Abhängigkeit von den Arbeitsbereichen und den zeitlichen Gegebenheiten gesehen werden muss.
Wichtige Rollentypen für das Wörterbuchprojekt sind:[1] Kanzlistin (DINAMLEX;1914-1942;nw);Gewährsperson(DINAMLEX;1913-lfd.;nw.|w.), Sammlerin (DINAMLEX; 1913-lfd.; nw.|w.); Praktikantin (DINAMLEX; 2009-lfd.; w.); Datenbankmitarbeiterin (DINAMLEX; 1993-lfd.; nw.|w.); Datenbankentwicklerin (DINAMLEX; 1993-lfd.; w.); Datenbankleiterin (DINAMLEX; 1993-lfd.; w.); Artikelverfasserin (DINAMLEX; 1963-lfd.; w.); Technische Redaktion (DINAMLEX; 1998-lfd.; w.); Redakteurin WBÖ-Online Edition (DINAMLEX; 2004-lfd.; w.); WBÖ-Gesamtredaktion (DINAMLEX; 1969-lfd.; w.); Stellvertretende Direktorin (2000-2005; w.); Direktorin(DINAMLEX;1998-lfd.,w.);Mitglied(Kuratorium;1994-2001;w.); Obfrau (SBT; 2006-lfd.; w.); Beiratsmitglied (SBT; 2006-lfd.; w.); Sprecherin des Beirats (SBT; 2006-lfd.; w.); Aktuarin der philosophisch-historischen Klasse (ÖAW; 1946-lfd.; w.); Mitglied der Gelehrtengesellsschaft (ÖAW; 1948-lfd.; w.); Vorsitzende der philosophisch-historischen Klasse (ÖAW; 2009-lfd.; w.); Vizepräsidentin (ÖAW; 2009-lfd.; w.).
Frauen spielten im Grundlagenbereich (Sammlung) u.a. auch zeitbedingt (frühes 20. Jahrhundert) eine untergeordnete Rolle. Nur 6% der ersten Sammlungen (und rund 90.000 Belege) gehen – wenn auch über das gesamte Bearbeitungsgebiet des WBÖ verteilt – auf Frauen zurück. Im Bereich der Archivierung und Sichtung des Materials spielten Frauen eine zentrale Rolle. Ihre aktive Mitarbeit im Schatten des Großprojekts hat von der Materialordnung und Materialerweiterung (Exzerption) bis hin zur Digitalisierung eine entscheidende qualitative und quantitative Rolle für das heutige Korpus. Die Digitalisierung steht von Beginn an bis heute unter weiblicher Konzeption und maßgeblicher Mitarbeit von

Frauen. Die Wörterbuchartikel werden seit den 70-er Jahren verstärkt von weiblichen Mitarbeiterinnen verfasst. Das derzeitige WBÖ-Team steht unter weiblicher Teamleitung. Bis heute spielen Frauen eine tragende Rolle und prägen das Bild des Projekts und Unternehmens entscheidend mit. Die Einträge in der Zeitspalte verdeutlichen auf anschauliche Weise, wie es gerade in den letzten Jahren vermehrt gelungen ist, Frauen als Entscheidungsträgerinnen zu etablieren.

1 Die für das Wörterbuchprojekt quantitativ und qualitativ wichtigen Rollentypen sind verzeichnet. In der Klammer angegeben sind: Organisatorischer Rahmen: Institut oder Akademieebene, Zeitraum, in der der Rollentyp wesentlich ist, wissenschaftliche (w) oder wissenschaftlich-unterstützende (wu) Tätigkeit. DINAMLEX = Institut für Österreichische Dialekt- und Namenlexika; SBT = Zentrum Sprachwissenschaften, Bild- und Tondokumentation; ÖAW = Österreichische Akademie der Wissenschaften.

> Dictionary, lexicon, glossary, wordbook or thesaurus? The usefulness of OALDCE7 and OLT for choosing the right word

DZIEMIANKO, ANNA

7 – Dictionary Use

Monolingual English learners' dictionaries (MLDs) published in recent years have many features which make them better suited to the needs of the target user group. Among others, onomasiology has slipped into their design. Today, MLDs typically list synonyms and antonyms, or even offer synonym notes, where words close in meaning are compared and contrasted. On the other hand, thesauri have also changed. The year 2008 witnessed the publication of the *Oxford Learner's Thesaurus: A Dictionary of Synonyms*, which goes beyond clustering words close in meaning. It defines each synonym, exemplifies its usage, and even juxtaposes selected synonyms in special notes.

The aim of the present study is to investigate the usefulness of the *Oxford Advanced Learner's Dictionary of Current English* (7th edition, OALDCE7) and the *Oxford Learner's Thesaurus* (OLT) for discriminating between synonyms. The paper is underpinned by empirical research, in which 73 advanced learners of English took part. In the experiment, words appropriate for given contexts had to be indicated in different synonym sets. The results reveal that neither dictionary significantly shortened the time needed to complete the task. Nonetheless, the use of OLT much more often resulted in successful synonym selection. Interestingly, synonym notes, present in both dictionaries, did not affect the subjects' choices. Besides, different information was usually referred to in the two dictionaries. In OALDCE7 the subjects paid attention most often to definitions, while in OLT – to

examples. The results of the supplementary questionnaire suggest that the students' familiarity with the two dictionary types could not have affected their performance. They were nonetheless more satisfied with their results when they had OLT at their disposal rather than OALDCE7. Yet, they were critical of the arrangement of synonyms in the OLT synonym clusters, where the alphabetical order, rather than frequency, would be a better solution.

> Donner un accès aisé aux formes phoniques des mots décrits dans un dictionnaire: étude pour un dictionnaire monolingue français destiné à de jeunes utilisateurs

GASIGLIA, NATHALIE

7 – Dictionary Use

Dans le cadre de cette contribution, je me propose de réfléchir à ce qui pourrait évoluer dans les dictionnaires sur support électronique concernant les descriptions des formes phoniques des unités linguistiques décrites et les modes d'accès à celles-ci. En envisageant les consultations de dictionnaires à la fois dans le cadre d'une aide à la compréhension (de ce qui est entendu ou difficile à déchiffrer) et à l'expression (énonciation ou lecture à haute voix, ou graphie des mots respectueuse de l'usage), je me propose d'examiner comment améliorer l'accès aux articles par les formes phoniques et l'utilisation des indications phonétiques fournies. Les orientations qui se dégagent de cette étude sont établies dans le cadre d'une création de dictionnaire électronique destiné à des élèves francophones de 11 à 15 ans (plus autonomes que les lecteurs débutants mais dont la maîtrise linguistique doit encore progresser) ou allophones de niveau intermédiaire ou avancé. Elles s'appuient sur ce qui est proposé dans trois dictionnaires publiés par Le Robert, l'éditeur français qui a attaché le plus de soin au traitement des prononciations dans ses produits: le *Petit Robert* électronique (éditions 2001 à 2008), qui est le plus élaboré des dictionnaires généraux électroniques français quant à l'accès aux descriptions des formes phoniques, le *Robert junior* électronique (éditions 1998 à 2006 – la dernière sous le titre *Le Robert des enfants*), qui dispose des mêmes fonctions de recherche que le premier, mais dont le texte, destiné aux élèves de 8 à 11 ans, est moins riche, et le *Robert oral-écrit* (1989), dictionnaire imprimé novateur pour apprenants (natifs ou allophones) qui permettait un accès aux graphies à partir des transcriptions de formes phoniques. Complémentairement aux modalités de traitement et de consultation inspirées de ces dictionnaires, le recours aux technologies de reconnaissance et de synthèse vocales est envisagé. Il impliquerait des partenariats de recherche et développement.

> **The ABBYY Lingvo Platform as a convenient tool for end users and a comprehensive solution for publishers**

KUZMINA, VERA

7 – Dictionary Use

Current paper is devoted to ABBYY Lingvo electronic dictionary, which is not only a dictionary software, but a family of software products, jointly termed the ABBYY Lingvo Platform. The ABBYY Lingvo Platform includes a range of dictionary software components, which are aimed at meeting same dictionary users' needs – finding an appropriate translation to the word in the given context. The software parts of ABBYY Lingvo Platform are available in different technological realization and for different usage scenarios: as a mobile software, a software for desktop computers or laptops, and as a client-server software made both for professional translators and language learning beginners. The components of ABBYY Lingvo Platform are: ABBYY Lingvo Desktop, ABBYY Lingvo Mobile, ABBYY Lingvo Server, and ABBYY Lingvo Content dictionary writing system. These products operate as well as in connection with each other through ABBYY data centre and independently from each other, for example when the internet connection is unavailable. The applications enable quick and easy access to the dictionary and reference content provided by dictionary publishers, other content providers and also to the content created by the dictionary users themselves. The main needs of dictionary users are covered, such as the ability to use dictionary content stored on different media (online, desktop and mobile), convenient tools for its usage – wide range of search and lookup capabilities together with user-friendly interface, and the possibility to use dictionary content from different content providers in one format and in the same software 'shell'. We believe that the ABBYY Lingvo Platform will cover the main market needs and help end users to get answer to their questions and give new ideas to content providers how to manage the content.

> **From Language-Oriented to User-Oriented Electronic LSP Dictionaries: A Case Study of an English Dictionary of Finance for Indonesian Students**

KWARY, DENY

7 – Dictionary Use

The rapid development of Internet technology and the significant increase in the number of non-native English speaking college students urge lexicographers working on LSP dictionaries to create better electronic dictionaries to satisfy the needs of these dictionary users. This paper argues that better LSP dictionaries can only be created if lexicographers move

from language-oriented to user-oriented lexicographical solutions. This paper shows that the traditional divisions of monolingual, bilingual and semi-bilingual dictionaries have confined the creation of lexicographical solutions that can thoroughly satisfy the needs of the users. The definitions given in monolingual LSP dictionaries are incomprehensible due to the use of difficult vocabulary. The equivalents given in bilingual dictionaries, though considered the quickest way for second language users to know the meaning of a term, do not really help the users when the equivalents relate to different concepts in L1 from the L2 due to cultural differences and when the equivalents are only the transfer of the L2 words. Combining the definitions and the equivalents, as in semi-bilingual dictionaries, may not work well either due to the overload of information presented to the users. Consequently, the shift from language-oriented to user-oriented has to take place in order to produce better lexicographical solutions. Better considerations on users' competences and characteristics are required in creating better electronic LSP dictionaries. In this paper, the implementation of this user orientation is only shown in the on-going project of an English dictionary of finance intended to give help to Indonesian college students to understand financial texts, but the proposed solutions may also be applicable to other LSP dictionaries with a similar type of users.

> **Users Take Shortcuts: Navigating Dictionary Entries**

LEW, ROBERT

7 – Dictionary Use

In the present paper we compare the effectiveness of two alternative meaning access facilitators in a monolingual learner's dictionary: a Menu system, placed at the top of a monolingual entry; and a Shortcuts system, where the cues are distributed throughout the entry. We test the two entry formats on 90 Polish learners of English at two CEFR levels, A2 and B1. The task which triggers dictionary consultation is guided translation from English to Polish. Three outcome measures are evaluated: access time to sense, accuracy of sense selection, and translation accuracy. While Menus and Shortcuts turned up no difference in terms of consultation speed, the task success was significantly better in the Shortcuts condition. Sense selection accuracy was also better, though not significanly so, for the Shortcuts. The overall conclusion of our study is that Shortcuts are more user-friendly than Menus, although this may also depend on the form of the cues and the medium of presentation.

> **One, Two, Many: Customization and User Profiles in Internet Dictionaries**

TRAP-JENSEN, LARS

7 – Dictionary Use

Recent textbooks in lexicography recommend the use of customization in e-dictionaries whereby users or dictionary-makers specify which information categories should be shown on the screen. In this paper I take a look at some online dictionaries and analyze how they solve the task. A few basic types are recognized, based on the answer to questions such as: are the user profiles specified by the user or by the lexicographer? Is the profile defined in relation to the look-up situation or to the user's general background and skills? Is the profile fixed or flexible? Must the profile be specified once and for all, before every look-up situation or can it be changed as the user navigates through the dictionary entry? For practical reasons, I confine myself primarily to English and Scandinavian dictionaries.
The analysis formed part of the preparatory phase of the online version of The Danish Dictionary. Four months after the introduction we can now observe from the log files how users manage the various options they are given. The experience so far is that user profiles that require deliberate action from the user are rarely used. The same holds for other kinds of customization such as advanced search possibilities. For the dictionary-maker there is all the more reason to be careful about configuring the default setting.

> **Monitoring Dictionary Use in the Electronic Age**

VERLINDE, SERGE AND BINON, JEAN

7 – Dictionary Use

The way in which a user consults a dictionary, navigates through a dictionary article and finds an answer to specific questions is a popular area of research in metalexicography. The successful development of online dictionaries opens new prospects in this area of research. Log files of online dictionaries may provide interesting 'free implicit feedback' (de Schryver and Joffe 2004:187). Thanks to its task- and problem-oriented interface, the *Base lexicale du français* (BLF) allows us to track all dictionary users' actions in a natural setting, outside any controlled research environment. Using these data, it should be possible to make well justified decisions on dictionary design.

> **O uso de dicionários na compreensão escrita em italiano LE**

ZUCCHI, ANGELA M.T.

7 – Dictionary Use

The use of dictionaries by FL students is a standard fact, but this use

is frequently questioned by FL teachers. Based on lexicological and lexicographic studies, in addition to language teaching, the survey described was developed in which empirical research is used as a way to answer the question of whether the help of a dictionary leads to differences in the success of understanding pre-determined lexical units, or whether the context itself is sufficient for such comprehension. To achieve this goal, we invited volunteer students enrolled in the undergraduate program in Italian as foreign language at FFLCH, USP, Brazil to participate in the survey. We established three groups of volunteers, the first using an Italian monolingual dictionary; the second using an Italian – Portuguese bilingual dictionary, and the third not using any dictionaries at all. The test consisted of four reading texts in which forty lexical units were highlighted, and whose proper meanings were to be verified by a multiple-choice test. After being collated, the results were submitted to a statistical analysis carried out by the Center of Applied Statistics (CEA-IME, USP). In addition to the statistical results, the methodology allows, through a template, the various elements present in both macro and microstructures of the dictionaries, which really helped comprehension according to the students, to be examined. The results of this empirical research demonstrated the important role of the dictionary in comprehending lexical units, and therefore, also in the teaching and learning process of a foreign language. Furthermore, this survey supported the study of Pedagogical Lexicography and teaching of the use of dictionaries in FL classes.

> The Treatment of Lexical Collocations of Six Adjectives Related to Feelings in A Sample of Bilingual Dictionaries English-Italian

BERTI, BARBARA

8 – Phraseology and Collocation

The importance of lexical collocations is nowadays undeniable, especially from a SLA perspective. Besides grammar, learners of a second language also need to access information concerning the lexical environment of words. The knowledge of the restrictions on lexical combinability is part of a L2 competence and should be successfully master by learners. Possibly, the most important source of information on lexis is the dictionary, thus it should also be (or become) a reference tool for the retrieval of collocations. In particular, bilingual dictionaries seem to be favoured by learners; this makes it increasingly more important for them to represent the combinatorial rules that organise lexis on the syntagmatic axis.
This study aims at investigating the presence and the treatment of some adjectival collocations in three bilingual dictionaries English-Italian. Six

adjectives related to the semantic area of emotions are singled out and their nominal, adverbial and verbal collocates looked up in both sections of the dictionaries. The data are then compared to those available from a dictionary of English collocations and from the British National Corpus. From a quantitative point of view, the study highlights an unsatisfactory presence of adjectival collocations, especially when the collocate is an adverb. The very few collocations found in the dictionaries are often closer to free combinations, therefore doubtfully useful to dictionary users. The collocations are not systematically organised and can be found interchangeably under either the base or the collocator, thus creating a feeling of confusion that leads to poor user-friendliness. On the whole, from the analysis of the data, it emerges that the issue of collocation should be taken into greater account from bilingual lexicography.

> **Onomasiologisch angeordnete Idiomlexika und ihr Nutzwert für die Translatologie: das Forschungsprojekt FRASESPAL zur deutsch-spanischen Phraseologie**

BUJÁN OTERO, PATRICIA

8 – Phraseology and Collocation

Die vorliegende Arbeit basiert auf einer praxisnahen und funktionalen Perspektive auf das Übersetzen von Phrasemen. Häufig werden diese als besonders problematisch beim Übersetzen empfunden. Für die intralinguale Äquivalenzanalyse ist aber erstens eine Unterscheidung zwischen Phrasemen im Text und Phrasemen im Sprachsystem nötig. Bei der Äquivalenzerstellung im ersten Fall müssen die Funktionen des Phrasems und deren Relevanz je nach Kontext und Kotext analysiert werden. Eine Hierarchisierung dieser Funktionen im Ausgangstext im Zusammenspiel mit den spezifischen Anforderungen des jeweiligen Übersetzungsauftrags und – vor allem – der Translatfunktion bestimmt dann eine oder andere Übersetzungsstrategie. Ziel dieser Arbeit ist eine Analyse dieser Aspekte, sowie der Vorteile, die die Erstellung von onomasiologisch angeordneten zweisprachigen Wörterbüchern für die Übersetzungspraxis hervorbringt. Die Analysebasis bildet der onomasiologisch angeordnete Thesaurus von Phrasemen aus dem semantischen Feld LEBEN/TOD, die im Rahmen des interuniversitäres Projekts FRASESPAL zur kontrastiven Phraseologie Deutsch-Spanisch zusammengestellt wurde. Idiomatik-Thesauri wie dieser eignen sich in besonderer Weise zu einem aktiven Gebrauch seitens des Wörterbuchbenutzers, was sich positiv auf die Übersetzungspraxis, sowie auf andere Bereiche, wie zum Beispiel die Didaktik einer Fremdsprache, bewirkt.

> **A proposal for an electronic dictionary of Italian collocations highlighting lexical prototypicality and the syntactic-semantic relations between collocation partners**

GIACOMINI, LAURA

8 – Phraseology and Collocation

My paper presents a corpus-based case study aimed at designing an electronic dictionary of Italian collocations and focussing on a small set of nouns belonging to the semantic field of *paura/fear*. In the paper, all of the steps involved in data retrieval, automatic and non-automatic evaluation, collocation selection and lexicographic organization are explained in detail. Lexicographic data are represented as a three-dimensional lexical framework displaying ontological, semantic and syntactic relations among lexemes. On the ontological level, *paura* as an entity is connected to contiguous emotions, but it also serves as a prototype for the category of the lexemes selected, shaping their syntactic and semantic behaviour. Collocations are formally categorized through a set of analytic parameters which enable a detailed lexical description as well as more finely grained dictionary search results. On the microstructural level, substantival collocation partners of the selected nouns are described in terms of thematic roles and semantic features, whereas adjectival collocation partners additionally have a set of principles derived from psychological studies applied to them. Finally, analysis of verbal collocation partners focusses on the interplay of the grammatical function and the thematic role of the noun they cooccur with, and on verbal Aktionsart.
The intended exicographic description rests upon a coherent cross-reference network linking the prototype to the other lemmas, collocation partners to each other, and collocations belonging to the narrower semantic field of *paura*, as well as collocations belonging to other semantic fields (e.g., the semantic field of emotions). At the same time, according to an open source principle, corpus-based lexical data can be deductively expanded using framework information.

> **Die Festlegung der Polysemie in einem phraseologischen Wörterbuch Spanisch-Deutsch**

HENK, ELISABETH AND TORRENT-LENZEN, AINA

8 – Phraseology and Collocation

Seit ungefähr sieben Jahren ist ein Team von Linguisten, Übersetzern und Studierenden mit der Aufgabe beschäftigt, ein Spanisch-Deutsches Wörterbuch der Redewendungen des europäischen Spanischs zu erstellen. Inzwischen haben wir in mehreren Publikationen die von uns

angewandten Kriterien bezüglich unterschiedlicher Aspekte wie zum Beispiel der Strukturierung des Definiens oder der Angaben über den eventuellen ironischen Gebrauch bekannt gemacht. Ziel unseres jetzigen Beitrages ist es, die Richtlinien zu erläutern, denen wir bei der Festlegung der Polysemie folgen.
Wir gehen in dieser Studie empirisch-induktiv vor und stellen unterschiedliche Probleme bei der Festlegung der Polysemie dar, die sich in unserer phraseographischen Praxis ergeben und die uns allmählich zu neuen Erkenntnissen führen. Wir vertreten die Meinung, dass das Wesentliche in einem phraseologischen Wörterbuch das Erfassen der übertragenen, phraseologischen Bedeutung ist und dass die Phraseographie aus diesem Grund eventuell andere Kriterien bei der Festlegung der Polysemie braucht, als diejenigen, die bei der monolexemischen Lexikographie gelten. Die Formulierung nützlicher Kriterien für die Festlegung der Polysemie in der phraseographischen Arbeit soll das Ziel einer anderen Publikation sein. Genauso wie es in der allgemeinen Lexikographie der Fall ist, gilt hier zu sagen, dass das Phänomen der phraseologischen Polysemie im Allgemeinen nicht durch streng systematische Kriterien erfasst werden kann. Gerade im Bereich der Phraseographie ist es jedoch wichtig, neue Technologien wie das Internet zu nutzen und die Methoden der linguistischen Pragmatik weiter zu entwickeln, damit nach und nach intuitive Vorgehensweisen durch wissenschaftliche Erkenntnisse ersetzt werden können.

> **Von 'hinkenden' Stühlen, 'tanzenden' Zähnen und 'verlorenen' Verkehrsmitteln.**
Erfassung und Darstellung italienischer lexikalischer Kollokationen für deutschsprachige L2-Lerner (auf der Grundlage des *Dizionario di base della lingua italiana – DIB*)

KONECNY, CHRISTINE

8 – Phraseology and Collocation

Im vorliegenden Beitrag soll ein am Institut für Romanistik der Universität Innsbruck geplantes Forschungsprojekt (Projektleitung: Ch. Konecny) vorgestellt werden, dessen Ziel in der Erfassung und Darstellung italienischer lexikalischer Kollokationen im Vergleich mit ihren deutschen Äquivalenten besteht. Die Kollokationsglieder selbst sind dem italienischen Basiswortschatz entnommen, so wie er im *Dizionario di base della lingua italiana – DIB* von Tullio de Mauro und Gian Giuseppe Moroni festgehalten ist. Kollokationen sind besondere syntagmatische Wortverbindungen, die von Muttersprachlern meist intuitiv korrekt gelernt und verwendet werden, für L2-Lerner hingegen eine häufige

Fehlerquelle darstellen, weil sie oft von jenen der eigenen Muttersprache abweichen. Ein Italienisch-Lernender sollte z.B. wissen, dass in dieser Sprache ein wackelnder Stuhl 'hinkt' (*la sedia zoppica*), ein wackelnder Zahn hingegen 'tanzt' (*il dente balla*) oder man einen verpassten Zug oder Bus 'verliert' (*perdere il treno / l'autobus*). Dem Projekt liegt ein enges Kollokationsverständnis zu Grunde, demzufolge lexikalische Kollokationen hierarchisch organisierte binäre Konstruktionen repräsentieren, welche aus einem kognitiv übergeordneten Element (Basis) und einem kognitiv untergeordneten Element (Kollokator) bestehen: In *la sedia zoppica* und *perdere il treno* z.B. sind *sedia* und *treno* die Basen und *zoppica* und *perdere* die Kollokatoren. Im Projektbericht soll anhand konkreter Beispiele u.a. gezeigt werden, wie die lexikographische Darstellung metaphorische Weiterentwicklungen und damit den polysemen Charakter von Kollokationsgliedern (*piantare un chiodo nel muro – piantare un coltello nella schiena di qcn.*), antonyme (*un coltello affilato – un coltello smussato*) oder alternative Komponenten (*levare / estrarre / cavare / strappare un dente*) und schließlich die jeweiligen referentiellen deutschen Äquivalente berücksichtigen kann. Die geplante Arbeit richtet sich an Italienisch- und Deutsch-Lernende, Italienisch- und Deutsch-Lehrende, aber auch an ÜbersetzerInnen und DolmetscherInnen, ist also sowohl als Lern- wie auch als Lehrhilfe gedacht. Darüber hinaus soll sie dazu beitragen, ein Bewusstsein für die Bedeutung von sprachspezifischen Kollokationen für das Sprachenlernen auch im Bereich der Lernerlexikographie Italienisch-Deutsch zu schaffen.

> Football Phraseology: A Bilingual Corpus-Driven study

MATUDA, SABRINA

8 – Phraseology and Collocation

Football is the most popular sport in the present century. It has assumed a role beyond that of a national sport by becoming a cultural manifestation. It is now a battleground for several important issues such as economy, the resolution of conflicts, poverty as well as racial and minority awareness (Anchimbe 2008). Football relations across countries have increased significantly over the past decades. In order to regulate these relations, we need to express ourselves through language, that language being, in most cases, English. However, each culture has its own way of playing and supporting its teams, a way that is differently expressed according to each mother tongue . The problem arises when there is an urge to express these particularities in a foreign language.

Football constitutes first and foremost a technical domain, though usually not considered as such. Therefore it involves a specialized language. Aiming

at understanding the football vocabulary we propose a detailed study of football phraseology. In order to do so the study is based on the notions of Corpus Linguistics (Bowker & Pearson, 2002; Hunston, 2002; Sardinha, 2004; Sinclair, 1991); *corpus-driven* translation studies (Tognini-Boneli, 2001) and terminology (Krieger & Finatto, 2004; Maia, 2002; Temmerman, 2000).
The study relies on the assumption that a term is not likely to be used apart from other lexical items and also on the fact that the protected status that is often attributed to it changes according to the context and the words to which it co-occurs.
The *corpus*, still being compiled, consists of approximately 285 thousand words – 156,146 in English and 127,984 in Portuguese.
Due to the complexity of compiling a representative corpus just a preliminary account of the findings is presented.

> **Coals to Newcastle or glittering gold? Which idioms need to be included in an English learner's dictionary in Australia?**

MILLER, JULIA

8 – Phraseology and Collocation

English idioms and figurative expressions are used by native English speakers of all ages and from many different English speaking countries. The non-literal nature of idioms can pose a problem for non-native speakers, however, who wonder why taking coals to Newcastle should be a significant action, or where the back of Bourke might possibly be. Many non-native speakers of English in Australia are university-age students, aged between 16 and 22, whose first point of departure in finding the meaning of unfamiliar expressions is likely to be a monolingual English learner's dictionary (MELD). Since the MELDs available in Australia are mainly of British origin, learners of English may therefore not find in them the Australian expressions that are used in general conversation and in the media. Moreover, Australian native speakers of English who belong to different generations may not know or use the same idioms. Students who do learn the meaning of an idiom need to know with whom it is appropriate to use such an expression, and this information is often not available in a MELD.
This paper addresses five idioms and expressions taken from a larger study of 84 idioms in order to examine which of these expressions are known and used by different age groups in Australia and the UK. Native English speakers in Australia and the UK completed 2085 surveys indicating where they had first encountered the 84 idioms and where they would use them. The findings indicate that not all expressions given in the British MELDs are known and used by native speakers in the 16-22 age range in either

the UK or Australia, and that Australians use idioms which are often not included in the British MELDs. It is therefore suggested that MELDs used in Australia include more Australian material, perhaps online or via a CD-ROM, and that a more appropriate labelling system be introduced to indicate age as a factor in usage.

> Part-Of-Speech Labelling and the Retrieval of Phraseological units

VRBINC, ALENKA

8 – Phraseology and Collocation

The paper presents some insights into the problems of PoS labelling of lemmata, paying special attention to locating phraseological units where it is essential to identify the correct part of speech of the word under which the phraseological unit is included and dealt with in monolingual learners' dictionaries. When studying the inclusion of individual words, senses and phraseological units in five learners' dictionaries (OALD7, LDOCE5, COBUILD5, CALD3, MED2), it was found that numerous lemmata are equipped with more than one PoS label. Consequently, the user no longer needs to identify each and every part of speech of the word in question. As far as the inclusion of phraseological units in the entry for a specific part of speech is concerned, another method of inclusion is proposed, which can be regarded as a further simplification of the microstructure: i.e., that all phraseological units with one common element belonging to different parts of speech are simply grouped together in one special idioms section without distinction between individual parts of speech. This method is certainly worth applying in monolingual learners' dictionaries.

> Phraseological false friends in English and Slovene and the metaphors behind them

VRBINC, MARJETA

8 – Phraseology and Collocation

The interest in false lexical equivalence reflects the interest in language contact, the observation of which always leads to the conclusion that formally identical and similar words and word combinations in different languages do not necessarily overlap semantically. Dictionaries of false friends deal with one-word lexical items, but false-friend relationship can also be established in phraseology. The aim of this paper is to look at phraseological components of English and Slovene lexicons with a view to identifying and describing the false semantic equivalence between idioms in these two languages.

When studying false lexical equivalence, the closeness or sameness of

form has been made *tertium comparationis*. Several phraseological units that are the same or similar in form but different in meaning in English and Slovene are analysed in the paper. Some of these pairs of idioms show certain common features, such as comparison, emotion, spoken or written communication. Phraseological false friends are illustrated by examples and similarities and differences between the idiom in English and the phraseological false friend in Slovene are commented upon.
Since phraseological as well as lexical false friends represent a great problem in communication, translation and lexicographic treatment, it is necessary to first raise awareness of the lexical traps into which non-native speakers of English as well as any other language may easily fall, regardless of their level of linguistic knowledge. It is, therefore, essential to find and treat these pairs of idioms appropriately and acquaint learners with them by including them in coursebooks, in bilingual, general and especially phraseological dictionaries.

> **Going organic: Building an experimental bottom-up dictionary of verbs in science**

WILLIAMS, GEOFFREY AND MILLON, CHRYSTEL

8 – Phraseology and Collocation

Choosing what headwords to enter in a dictionary has always been a major question in lexicographical practice. Corpora have greatly helped ease both the choice of words to add, and those to remove, by resorting to frequency counts so as to monitor usage over time. This has been particular valuable in the building of learners dictionaries as, however good earlier word lists may have been, they were built largely in intuition whereas, corpora allow the consultation of large reference corpora for a better picture of current realities. In specialised dictionaries dealing with terminological issues, pure frequency is not a feasible solution for headword extraction. However, linked with extraction patterns and statistical tools, corpora still play a major role in supplying information on terms in use. In this research we aim to tackle a situation that lies in between the needs of an advanced learners dictionary and those of a terminological dictionary in attempting to build a pattern dictionary for verbs used in scientific research papers. In order to select verbs for this dictionary and put them into classes, we propose to use collocational relationships as a tool for both selection and analysis of patterns. The principle here is that a series of high frequency verbs can provide the seeds from which prototypical patterns can be extracted. By moving backwards and forwards from verb to argument and back pattern are revealed that use the statistical selectionning to highlight verbs lower in the frequency

list that would otherwise be overlooked. Thus patterns will naturally enlarge the word list by selecting what is statistically significant with a textual environment. These patterns not only illustrate typical usage in a specialised environment, but will also group verbs according to textual functions as authorial positioning and description of processes.

> Sampling techniques in metalexicographic research

BUKOWSKA, AGNIESZKA ANUSZKA

9 – Lexicological Issues of Lexicographical Relevance

Browsing through International Journal of Lexicography archives and other metalexicographic work one could easily notice that sampling techniques are generally neglected by metalexicographers, rarely described exhaustively by the authors themselves and almost never discussed, even though numerous researchers sample in order to make generalizations about the whole dictionary text, usually too large to be studied in its entirety. Not rarely samples consisting of one stretch only, usually selected judgmentally, are used to draw inferences about the whole dictionary text and serve as a basis for statistical analysis, which produces results of uncontrolled reliability. This study aims both at exposing the pitfalls of currently used sampling techniques and at proposing probability sampling instead.
Two basic probability sampling schemes were examined: simple random and stratified selection of pages. Censuses based on three dictionaries, three characteristics examined in each one, confirmed my concerns regarding one-stretch sampling. Simple random selection of pages produced, as expected, far more satisfying results in virtually all the cases. This can be, however, bettered by stratification in case of entry-based characteristics in larger dictionaries. Page-based characteristic, mean number of entries per page in this study, did not benefit from stratification. The smallest of my dictionaries presented a range of problems mostly connected with stratified sampling. Furthermore, empirical evaluation of sampling techniques proposed in Coleman – Ogilvie (2009) demonstrated that randomization within strata is also crucial.

> 'Offensive' items, and less offensive alternatives, in English monolingual learners' dictionaries

COFFEY, STEPHEN

9 – Lexicological Issues of Lexicographical Relevance

This paper discusses lexical items which have been labelled as 'impolite', 'offensive' or 'rude' in monolingual learners' dictionaries (MLDs). Such items may be grouped into three broad categories. Firstly, there is lexis

which relates to the human body and its functions (e.g. *knockers, dick, to crap, to screw*). Secondly, there are items which refer to people and which are potentially insulting (e.g. *bitch, dago, midget, queer*). Thirdly, there are words and phrases, with a variety of meanings, which have in common the fact that they make use of the potentially rude words referring to the human body. Examples are *to balls something up, not to give a shit, fucking, a piss artist* and *work your arse off*.

The precise aim of the paper is to draw attention to the fact that, wherever possible, learners should be provided with less offensive alternatives to the potentially offensive lexis.

In order to assess the current situation in MLDs, a study was carried out on over 200 such lexical items in recent editions of five dictionaries. The main conclusions reached were that in many cases learners are not being provided with alternative lexis, or else that the alternatives suggested are somewhat banal in nature. It is also proposed that in some cases a contextualized example of a lexical item could be rewritten in order to show learners what a less offensive version of the example would look like.

> **Deiktische Konstruktionen des Deutschen aus lexikographischer Perspektive**

DOBROVOL'SKIJ, D.O.

9 – Lexicological Issues of Lexicographical Relevance

In German, there are two quasi symmetrical constructions: *vor sich hin* and *vor sich her*. These constructions are relatively frequent; however, their meanings have not been sufficiently investigated and their lexicographic description remains rather poor. These constructions are highly idiosyncratic, at least from the perspective of other languages, i.e. speakers have no chance to use them properly if they do not learn them as idiomatic units of the lexicon. The reason is that the meaning of these constructions does not come about as a result of the composition of meanings of their constituent parts. In this sense, these two constructions have also to be studied within both phraseology and Construction Grammar.

The best way to deal with the semantics of *vor sich hin* lexicographically is to postulate a prototypical meaning of this construction and to impose semantic rules (in the sense of Apresjan), which modify the prototypical meaning according to the context. In other words, here we are dealing with coercion. The semantic features that are good candidates for the structure of the prototypical meaning are 'duration', 'introversion', 'week intensity', 'uncontrallability', 'not result-oriented'. In every single VP-construction, the prototypical meaning is specified through focusing some of the semantic features and/or deleting others.

The corpus-based analysis of the construction *vor sich her* showed similar results. From the theoretical perspective, this construction deserves special attention because of the semantic contribution of the deictic element *her*. It is obvious that here we are not dealing with the 'standard meaning' of *her*. The deictic element *her* focuses here the idea of the 'parallel movement'. Obviously, the deictic elements *hin* and *her* have a much richer semantic potential than is assumed in the traditional lexicological and lexicographic description.

> **Onomasiological dictionaries and ontologies**

FRANÇA, PATRÍCIA CUNHA

9 – Lexicological Issues of Lexicographical Relevance

There has been an increasing interest in ontologies over the last years by the computer science. This interest can be attributed to the advent of what has been known as Semantic Web. Consequently, a persistent interest in Onomasiological Llexicography has arisen and several authors have written about the relation between thesaurus and ontologies, determined to build bridges between the two representational instruments.
The aim of this paper is to light up the discussion between the similarities and differences between onomasiological dictionaries and ontologies. It also presents definitions for concepts such as onomasiology, onomasiological dictionaries, ontologies, formal ontologies, linguistic ontologies.
It intends to demonstrate that the differences pointed out by some authors to distinguish onomasiological dictionaries from ontologies are not quite evident.

> **Terminology, Phraseology, and Lexicography**

HANKS, PATRICK

9 – Lexicological Issues of Lexicographical Relevance

This paper explores two aspects of word use and word meaning in terms of Sinclair's (1991, 1998) distinction between the *open-choice principle* (or *terminological tendency*) and the *idiom principle* (or *phraseological tendency*). Technical terms such as strobilation are rare, highly domain-specific, and of little phraseological interest, although the texts in which such word occur do tend to contain interesting clusters of domain-specific terminology. At the other extreme, it is impossible to know the meaning of ordinary common words such as the verb blow without knowing the phraseological context in which the word is used.
Many words have both a terminological tendency and a phraseological tendency. In some cases the two tendencies are in harmony; in other

cases there is tension between them. The relationship between these two tendencies is investigated, using examples from the British National Corpus.

> **An outline for a semantic categorization of adjectives**

HEYVAERT, FRANS

9 – Lexicological Issues of Lexicographical Relevance

The aim of this paper is to sketch some basic principles on which a full semantic categorisation of adjectives can be founded that will allow for constructing uniform description templates for the individual categories. The underlying idea is that there should be a one to one correspondence between the category and the template used, like it is the case for most of the existing categorisations of nouns. For that purpose adjectival categories need to be defined by 'adjectival' expressions (mostly with present or past participles as their head). Since this procedure generates only a limited number of very general categories (more or less parallel to the verb categories containing static relations), the templates are extended with some semantic features among which the feature *domain*, plays an important role since it offers the opportunity to specify on a subcategorial level the information that most existing proposals for adjective categorisation give: the conceptual field in which the adjective is to be situated. These conceptual fields for adjectives appear moreover to play an important part in the construction of noun templates, not in terms of form but also in terms of content. The ultimate aim of this proposal is to construct a kind of template building grammar the elements of which can be used equally for nouns, verbs and adjectives.

> **Inflection and Word Identity**

JANSSEN, MAARTEN

9 – Lexicological Issues of Lexicographical Relevance

This article argues that inflection should be seen as a (partial) criterion that defines homonymy: when two word meanings have different inflected forms, they have to belong to different lexical entries. If this were not the case, it could not be maintained that inflection is a property of lexical entries, but we have to rather say that each word sense has its own inflectional paradigm, even though in most cases all senses of a word inflect in the same way. Although there are apparent cases where it looks like inflection might be in fact dependent on word meaning, none of these cases really goes against the hypothesis that inflection is a property of lexical entries, and not of word senses.

> Defining Dictionary Definitions for EFL Dictionaries

KERNERMAN, ARI AND GEFEN, RAPHAEL

9 – Lexicological Issues of Lexicographical Relevance

The following are some of the issues involved in defining headwords, which I will touch on in this paper:
What is the definition of a dictionary definition?
What linguistic styles for definitions are in use today for specific types of dictionaries? (e.g., technical definitions, folk definitions)
Are there differences in style and technique for defining the various parts of speech?
Is it valid to explain a word in terms of a different part of speech for the sake of clarity?
Are definitions in corpus-based dictionaries different from those in non-corpus based dictionaries?
Can (and should) the viewpoint of the lexicographer be completely hidden?
Does saying what a word is not, adequately explain what it is?
Should only active voice be used?
Which is more important, accuracy or comprehensibility?
How do definitions in learners' dictionaries differ from those in general-purpose dictionaries?
Should the definition be translated for reinforcement, in FL learning?
How useful are illustrations?
How useful are synonyms? Can they replace definitions?

> Systematic Polysemy of Nouns and its Lexicographic Treatment in Estonian

LANGEMETS, MARGIT

9 – Lexicological Issues of Lexicographical Relevance

The focus of the study of the polysemy of the Estonian noun (Langemets 2010) was on identifying the systematic patterns of noun polysemy with further perspective to elaborate the principles to encode and represent systematic polysemy of nouns in the database of the one-volume dictionary of Estonian (to appear in 2015) and in the **≡≡Lex** (= EELex) dictionary management system of the Institute of the Estonian Language.
The analysis was based on the lexical perspective, i.e. on the lexicographic representation of polysemy in the academic six-volume monolingual dictionary of Estonian (1st ed. 1988–2007, 2nd ed. 2009), and the supportive theory of generative lexicon by means of a qualia structure (Pustejovsky 1995). The sample of study consisted of simple nouns (843 headwords in

all), the total of 1738 semantic units covered both the numbered senses and various subsenses. A hierarchy of the semantic types of nouns, adapted from the lexicographic projects SIMPLE[1] and CoreLex[2], as well as the Estonian Wordnet[3], was used as an ancillary means of analysis enabling, in a way, to 'measure' the regularity of alternating word senses.
A result of the analysis is the list of 40 systematically polysemous patterns, presented as the 'golden standard' of systematic polysemy in Estonian (named after Peters 2004). In total the sample (843 headwords) contained 305 sense alternations that could be interpreted as revealing systematic polysemy. Of those, nearly every fourth (72 patterns) involves an ARTEFACT sense, while half (!) of the patterns involve ACTIVITY.

1 SIMPLE homepage, see http://www.ub.es/gilcub/SIMPLE/simple.html (31.03.2010).
2 CoreLex Online, see http://www.cs.brandeis.edu/~paulb/CoreLex/corelex.html (31.03.2010).
3 Estonian Wordnet, see http://www.cl.ut.ee/ressursid/teksaurus/ (31.03.2010).

> Une pratique lexicographique émergente: les dictionnaires détournés

LÉTURGIE, ARNAUD

9 – Lexicological Issues of Lexicographical Relevance

Parmi les nombreuses références bien installées auprès du public, certains auteurs agrémentent le paysage lexicographique français de dictionnaires particuliers: les dictionnaires détournés. Composés de néologismes construits par des procédés tant morphologiques que sémantiques, ces dictionnaires soulèvent des questionnements qu'il est bon de considérer. L'émergence de plus en plus prégnante de dictionnaires de mots inventés conduit des perplexicologues (les 'linguistes hagards devant la prolifération des mots-valises' selon Finkielkraut) à observer ce type de production lexicographique.
L'objet de cette étude sera donc d'exposer les principes du détournement lexicographique afin de mettre en évidence l'apparition d'un genre lexicographique à part entière: les dictionnaires détournés. Nous en présenterons la typologie pour en illustrer la diversité. Ces ouvrages ne sont en effet pas tous construits à l'identique et trois catégories de dictionnaires se dégagent au sein d'un corpus de 40 références, selon les méthodes de création lexicale employées pour bâtir leurs nomenclatures.
Nous présenterons quelques données quantitatives reflétant la multitude d'ouvrages de ce genre parus entre 1979 et 2010. Cette analyse sera l'occasion d'observer la prééminence d'une catégorie, celle des dictionnaires de mots-valises, sur les deux autres. Enfin, il nous sera également permis d'aborder (de façon superficielle) les différents apports théoriques et pratiques du

détournement de dictionnaire. Bien qu'ils s'attachent à pasticher le modèle lexicographique classique, les dictionnaires détournés empruntent nécessairement des spécificités à leurs modèles. Ainsi, certains auteurs présentent leurs dictionnaires – de mots-valises essentiellement – comme des ouvrages didactiques permettant à des locuteurs étrangers ou aux enfants d'acquérir le vocabulaire du français. D'autres militent de façon parodique pour l'intégration de leurs néologismes dans le dictionnaire de l'Académie française.
Tous ces aspects font des dictionnaires détournés un vaste champ d'investigation que nous introduirons dans cet article.

> Multilexical units and headword status. A problematic issue in recent Italian lexicography

MARELLO, CARLA

9 – Lexicological Issues of Lexicographical Relevance

The paper will discuss the headword status of multilexical units in Italian monolingual dictionaries and will include a comparison of Italian and Spanish dictionaries. Twentieth century monolingual lexicographies of Romance languages recognized and registered multiword units, but did not promote them easily to headword status. Italian and Spanish monolingual lexicography in particular have very few multilexical units whereas French has a few more. The initial infiltrations through the 'one-word headword' wall came through Latin borrowings (*alter ego* 'second self', *aut aut*, 'forced choice', *tabula rasa* 'blank sheet'), through two (or more for French) centuries of French and Anglo-American multiword borrowings entering gradually into the Italian language and then into monolingual dictionaries macrostructures (for instance *ballon d'essai* 'trial balloon', *malgré lui,* 'despite him', *fair play, self-made man* are XIX century borrowings; *j'accuse*, 'denunciation', *au pair*, *best seller*, *on the road* are XX century borrowings), and in recent decades through the macrostructures of bilingual English-Italian dictionaries where English multilexical headwords are registered and brought to the attention of Italian monolingual lexicographers as multiword units with headword status in English monolingual dictionaries. A status which might determine them becoming multilexical headwords also in Italian monolingual dictionaries. Nowadays most Italian multiwords still remain registered under one-word headwords, even adjectival or adverbial phrases which cannot occur as single words (as for instance *alla carlona* 'carelessly', *a perdifiato* 'at the top of one's voice' registered under the headword **carlona, perdifiato,** words with combinatorial usage only. Italian corpora can help define the confines of the multilexical

unit and establish possible variations, such as widespread elliptical uses. Coherent corpus-based decisions are in turn extremely valuable not only for lexicographers, but also for POS tagging of corpora in which the multilexical units are recognized and entered as a whole in addition to the single parts.

> A Semantic and Lexical-Based Approach to the Lemmatisation of Idioms in Bilingual Italian-English Dictionaries

MULHALL, CHRIS

9 – Lexicological Issues of Lexicographical Relevance

The aim of this paper is to propose a new semantic and lexical-based lemmatisation framework for the recording of idioms in bilingual Italian-English dictionaries. Many of the difficulties and inconsistencies characterising the lexicographic treatment of idioms stem from them being viewed as a semantic and lexically homogenous phrasal category. This incorrect generalisation typically motivates the traditional description of idioms as being non-compositional and lexically fixed units. Current bilingual Italian-English dictionaries treat idioms quite unsystematically, mainly due to their reliance on the subjective judgement of lexicographers and generic syntax-based listing strategies. The rationale for pursuing these methods remains unclear, particularly given the availability of substantive semantic and lexical information that could provide a more defined template for determining the position of idioms in a dictionary. This paper looks at two particular aspects of idioms in five current bilingual Italian-English dictionaries: *Il Ragazzini* (ZIR) (2009), *Grande Hoepli Dizionario Inglese* (GHDI) (2007), *Il Sansoni Inglese* (SI) (2006), *Oxford Paravia Italian Dictionary* (OPID) (2006) and *Hazon Garzanti Inglese* (HGI) (2009). The first is a semantic-based investigation, which analyses the entry procedures for 150 English and 150 Italian idioms across three categories: pure idioms, figurative idioms and semi-idioms. The second examines the listing strategies for 40 English and 40 Italian idioms with variable verb and noun components. Overall, two particular trends emanate from the analysis. Firstly, the arrangement of idioms is unsystematic and the allocated entry points do not reflect or emphasise their individual semantic or lexical features, which are central to their identity. Secondly, the English-Italian and Italian-English sections of certain dictionaries are disparate in their overall coverage with Italian idioms assigned a greater number of listings. These discrepancies call for a formulaic entry model that eliminates the subjectivity, inconsistency and unsystematic approach currently associated with the treatment of idioms in bilingual Italian-English dictionaries.

> **Seeing through dictionaries: On defining basic colour terms in English, Japanese and Polish lexicography**

PAKUŁA, ŁUKASZ

9 – Lexicological Issues of Lexicographical Relevance

It seems that despite the undeniable fact that colour research has received considerable attention for centuries resulting in more than 3000 publications during the last 150 years (MacLaury 1997 after Steinvall 2002), there still exists a niche to be filled. There has been no or very little research regarding colour terms conducted from the viewpoint of (meta) lexicography. The present study is meant to evaluate existing dictionary definitions of Basic Colour Terms (henceforth BCTs) from the colour lexicon of English, Japanese and Polish in order to detect any doubtful content which could be improved to equip the dictionary user with richer, more adequate information regarding the colour lexicon. The immediate aims of the study are to determine: 1) what definition types are used to define CTs 2) what prototypes extensional definitions point to when defining BCTs and how these relate to the data obtained from naive native speakers of the languages in question. To this end, two empirical investigations were conducted. The first one is devoted to dictionary definitions, while the second one is an experiment carried out among naive native speakers of the three languages.

> **Getting through to phrasal verbs: A cognitive organization of phrasal verb entries in monolingual pedagogical dictionaries of English**

PERDEK, MAGDALENA

9 – Lexicological Issues of Lexicographical Relevance

The abundance of English phrasal verbs along with their syntactic and semantic complexity has always been a stumbling block for learners of English. Some think of phrasal verbs as hallmarks of a native-like command of English but there is no universal method to learn their natural contexts or applications and no ready-made recipe to deduce their meaning is available. Therefore, more attention should be paid to the accurate lexicographic description of phrasal verbs in learners' dictionaries, which are often the first source of reference for students. Moreover, dictionary compilers should aim at such presentation of these structures as to guide the users towards working out the multiple meanings of phrasal verbs on their own by creating cognitive links in the entries or even offering spatial cognitive networks.
The paper looks at the organization of a phrasal verb entry in the most recent pedagogical dictionaries of English from the cognitive perspective.

The layout of the entries is examined with focus on the methods used to differentiate the many meanings of phrasal verbs, especially figurative ones and an attempt is made to find any cognitive links that are used to generate helpful associations and predictions about the meaning. In his recent paper on phrasal verbs, Brodzinski (2009) calls for such an associative approach to presenting phrasal verbs to learners, be it in class or in a dictionary. His claim is that for pedagogical purposes it is better to replace the multiple meanings of a given phrasal verb with one core meaning along with applications.
An alternative to the linear organization of a phrasal entry could be a network of meanings underlying any possible cognitive links between different senses. Such an approach might prove to be more stimulating for non-native users. Three examples of such networks, each with different semantic focus, are presented in the paper.

> **The Frisian Language Database as a tool for semantic research**

SLOFSTRA, BOUKE AND VERSLOOT, ARJEN

9 – Lexicological Issues of Lexicographical Relevance

In this paper, the authors present two examples from the field of body parts, illustrating the level of details, very often with a diachronic component, that can be detected in the study of semantics. The bilingual background of Frisian – Dutch is in one way or another a second language in Friesland already since the late Middle Ages – constitutes an extra trigger in the organisation and restructuring of the lexemes and their meanings. The given examples from Google and the Frisian Language Data Base illustrate that a diachronic and comparative corpus based approach can add several aspects that have remained uncovered so far in the more traditionally conceptualised WFT. This enhanced picture of meanings and their developments are an essential prerequisite for gaining a deeper understanding of the organisation and structure of human semantic concepts and their operationalisation in human speech.

> **From lexicological to lexicographical issues: Italian verbs with predicative complement**

STRIK LIEVERS, FRANCESCA

9 – Lexicological Issues of Lexicographical Relevance

Intransitive (e.g., *become*) and transitive (e.g., *consider*) verbs obligatorily requiring a predicative complement are an interesting, and at the same time problematic issue both at a theoretical and at a lexicographical level. In this paper we focus on Italian verbs, and on the way two computational semantic lexica deal with them. Both in ItalWorNet and in SIMPLE the

treatment of these verbs shows to be problematic, since the information appears to refer to the 'verb + predicative complement' complex rather than to the verb itself. Recognizing that verb and predicative complement contribute to the construction of a unitary event, we believe that it is nevertheless possible, and useful, to isolate the role of the two components. The description proposed here is based on the Generative Lexicon model (Pustejovsky 1995), and it is in line with the recent project of a lexical resource for (sub)event structure (Im and Pustejovsky 2009). Verb and predicative complement codify each a different part of the subevent structure. To give an example, '*diventare* ('become') + predicative complement' is a transition, where *diventare* codifies the process subevent, and the predicative complement codifies the (result) state subevent. This kind of analysis can possibly be integrated into the SIMPLE lexicon, which is already built following the Generative Lexicon model.

> **On defining the category MONSTER – using definitional features, narrative categories and Idealized Cognitive Models (ICM's)**

SWANEPOEL, PIET

9 – Lexicological Issues of Lexicographical Relevance

This paper explores how the coherence between a lexical item which denotes a category and the lexical items that refer to individual members of the category can be expressed in explanatory dictionaries. A detailed analysis is provided of the relationship between the lexical item *monster* (which refers to a category) and the lexical items that refer to individual members of this category (e.g., Cyclops, dragon, mermaid, vampire, werewolf, Dracula, and zombie). More specifically, the goal of the paper is to determine whether the semantic explanation(s) for *monster* could function as a dictionary internal (as opposed to Fillmore's (2003) external) cognitive frame for the other lexical items in the monster set. If not, the question is whether and how the field of monsterology could assist one in designing such a frame and what the content, structure and function of such a frame would be.
In Section 2.1 the focus falls on current lexicographic practices and problems in defining the category monster and its members. The dictionary entries for *monster* and those of a number of its members in a selection of English explanatory dictionaries are surveyed to determine what cognitive models of the category monster underlie these definitions. In Section 2.2 the focus falls on the definitional features, ICM'S and narrative structures used to define the category of the monster in the field of monsterology and on the numerous meanings monsters may have as symbolic expressions (metaphors in particular). Section 3 shortly summarizes the contribution monsterology could make towards the definition of a monster frame.

> **Metonymical Object Changes in Dutch: Lexicographical choices and verb meaning**

SWEEP, JOSEFIEN

9 – Lexicological Issues of Lexicographical Relevance

The Dutch term *objectsverwisseling* (literally: 'object change') is a lexicographical label used to describe specific combinations of a verb with two qualitatively different direct objects. Illustrative examples are *de borden / de tafel afruimen* ('to clear the plates / the table'), *hout / een vuur / de haard aansteken* ('to light wood / a fire / the fireplace'), *riet / manden vlechten* ('to weave reed / baskets'), *gaten / sokken stoppen* ('to darn holes / stockings'), *sinaasappels / sap persen* ('to press oranges / juice'), *eieren / kuikens uitbroeden* ('to hatch eggs / chicks'), etc.

These examples are often analysed as specific instances of metonymy (cf. Adelung 1811; Van Dale 2005; Koch 2001; Waltereit 1998). Both possible direct objects are interchangeable because they are conceptually connected by their existence as a conceptual unity in the real world (such as a set table, a wood fire, reed baskets, etc.). There are, however, some discrepancies between linguistics studies of metonymical object changes (MOCs) and lexicographical choices in dictionaries. These basically concern the question of whether an object change affects the meaning of the verb.

On the basis of theoretical considerations as well as lexicographical descriptions I will try to clarify to what extent MOCs influence the meaning of a verb. To this purpose, I will evaluate the incorporation of MOCs in two standard Dutch dictionaries, i.e. *Van Dale Groot Woordenboek van de Nederlandse Taal* (2005) and *Woordenboek der Nederlandsche Taal*. Theoretically, it will turn out to be necessary to distinguish between grammatical-relational information and lexical meaning (cf. Brdar 2007: 181). I will argue that MOCs actually provide evidence for the fact that the verb has one lexical meaning. In this way, the present paper gives more insight into the object changes, into the underlying metonymy and also into verb meaning in general. These insights may subsequently be useful in the improvement of dictionary entries.

> **Metonymy representation in English monolingual learners' dictionaries: Problems and solutions**

WOJCIECHOWSKA, SYLWIA

9 – Lexicological Issues of Lexicographical Relevance

The paper aims to show how the tenets of the cognitive theory of metonymy can benefit the representation of metonymic lexemes in pedagogical lexicography, so that the semantic connections between

basic and derived meanings become more transparent and motivated. It reports the results of a lexicographic study into the representation of conventionalised metonymic lexemes in the five most renowned English monolingual learners' dictionaries (henceforth MLDs): CALD2, COBUILD4, LDOCE4, MEDAL2 and OALDCE7. The study focuses on three elements of the dictionary entry: sense arrangement, definition, and the correlation between noun codification and exemplification. These features are evaluated against the background of both the cognitive theory of metonymy and the widely accepted principles of lexicographic practice. Significant inconsistencies concerning the treatment of metonymy are found within each dictionary, as well as numerous cases where the semantic relationship between the source and target senses of a metonymic lexeme is broken. It is also noticed that in the case of metonymisation which results in change of noun's countability, noun codes are sometimes ambiguously assigned, and some examples of usage do not explicitly show the count-mass distinction. Solutions are offered to arrive at a more systematic, transparent and cognitively oriented representation of metonymy. These include using template entries in the compilation process, subsuming the definition of the metonymic target under the source definition, and defining the target as a semantic elaboration of the source.

> **The German-Lower Sorbian Online Dictionary**

BARTELS, HAUKE

10 – Lexicography of Lesser Used or Non-State Languages

After the publication of a new and comprehensive Lower Sorbian-German dictionary in 1999, the urgent need for an active learner's dictionary has been widely felt. Some specifics of the sociolinguistic situation of Lower Sorbian must have direct impact on the conception of such a dictionary: For almost all speakers of younger generation German is the first and better known language. German-Lower Sorbian interference, a very small or only partially elaborated vocabulary, and an often defective command of grammar, especially of those parts of it lacking in German, is widespread. Since 2001 the Lower Sorbian Department of the Sorbian Institute works on a dictionary that tries to meet the requirements of that target-group. With respect to the fact that Lower Sorbian is highly endangered and there is no time to lose, all information is published on the internet as quickly as possible. In 2003 a first version of the online dictionary 'Deutsch-niedersorbisches Wörterbuch' (DNW) was launched. At the present the DNW contains about 70,000 entries, but it will continually be extended and corrected; it is still considered a draft version.

Apart from some technical background information, the paper gives an overview of the lexicographic description. In order to help to avoid typical L1-interferences and to actively use the minority language the dictionary offers, for example, additional information about the use of verbal grammatical and lexical aspect (*Aktionsart*). Also support verb constructions (so-called *Funktionsverbgefüge* in German), where direct translations of the German construction often lead to a non-idiomatic language usage, are taken into consideration. For a better integration of such and other important information, some new conventions have been introduced, hoping that the DNW will function as a learners' dictionary as well as a contribution to language documentation.

> **WFT: The comprehensive Frisian Dictionary (*Wurdboek fan de Fryske taal / Woordenboek der Friese taal*)**

BOERSMA, PITER

10 – Lexicography of Lesser Used or Non-State Languages

The *Woordenboek der Friese taal* is a dictionary of a regional minority language. Yet it may be compared to the big scholarly dictionaries of national languages like Dutch, German and English, not because of its size but with respect to its principles. The *Woordenboek der Friese taal* is, as a description of a minority language, in this sense unique. Its more modest size is partly due to the dictionary's design, but a more important reason is that the lexicographical description of Frisian is hampered by the absence of a large variety of written sources, because Frisian, characteristically as a minority language, especially functions as a spoken language.

In my paper I clarify how the position of a minority language – and in addition the scholarly infrastructure – are decisive for the lexicography of Frisian and the compilation and contents of the *Woordenboek der Friese Taal* in particular.

Before discussing some aspects of the WFT itself I will deal with three items. 1. The unfinished dictionary *Lexicon Frisicum (A-Feer)* (1872) by J. H. Halbertsma (1789-1869), the founding father of the lexicography of modern Frisian; 2. The continuation of Halbertsma's lexicographical work, resulting in the *Friesch Woordenboek* (1900-1911). 3. The preamble to the *WFT*.

I discuss the following aspects:

- the choice of the non-Frisian metalanguage in the dictionaries above
- the choice of only post-1800 Frisian in the WFT.
- the choice of regional variants in *Friesch Woordenboek* and WFT
- the choice of including the first attestation of each entry into the WFT

– the microstructure of the WFT
– etymology in the WFT

I finally mention the future of the WFT: with the completion of the the paper dictionary, the WFT is now ready to enter the exiting world of online electronic dictionaries.

> **The language norm in a century of Frisian dictionaries**

DUIJFF, PIETER

10 – Lexicography of Lesser Used or Non-State Languages

Since the Renaissance in Western Europe language builders have been making efforts to standardise languages. In the Netherlands and in Belgium the Dutch standard language became accepted generally. The standardising process of Frisian, the second official language in the Netherlands, was different and it still is. The standard for Frisian had not crystallized out yet. In practice, this means that several variant forms and pronunciations are accepted. A series of Frisian dictionaries have been published in the past hundred years. In this contribution the question will be answered whether these dictionaries contributed to standardising Frisian. Did the dictionaries reflect dialectal diversity or did they have a prescriptive design? In answering this question the position of four phonological variations in the dictionaries has been investigated. Also has been made an inventory of the editor's comments on their selection of dialect forms.
On the base of the results the conclusion must be that Frisian dictionaries did not use one and the same standard language consequently. A small number of dictionaries consciously prefer to include only one variant entry form. Just the more elaborate and widely used dictionaries show a rather tolerant standard of language, though not in consequent fashion. The electronic language databases of the Fryske Akademy show that in practice the choices made by the most frequently consulted dictionaries were followed generally.
In the paper also a picture of the Frisian language and a briefly description of the history of Frisian lexicography have been given.

> **Can the new African Language dictionaries empower the African language speakers of South Africa or are they just a half-hearted implementation of language policies?**

KLEIN, JULIANE

10 – Lexicography of Lesser Used or Non-State Languages

Language planning was always a very sensitive topic in South Africa, as languages was used to separate people during apartheid. This presentation

analyses three different Sesotho sa Leboa dictionaries, which can be seen as examples of a successful implementation of language policies. The policies which are discussed here are the constitution of the Republic of South Africa form 1996, The National Lexicographic Units Bill from 1996 and the South African Languages Bill from 2000. The main objective of those language policies is the development and promotion of the eleven official South African languages. Dictionaries are one possibility to develop languages, .e. they describe the standardised variety of a language. They can be used as tools to promote the African languages, as they are the visible proof that the language has the words to be used in a specific situation, for example a dictionary of Maths shows that the language has words for mathematical concepts.

The three dictionaries which are discussed here are a Sesotho sa Leboa – English general dictionary which was published by the Sesotho sa Leboa National Lexicographic Unit, a bilingual Sesotho sa Leboa English school dictionary published by OUP South Africa and a Sesotho sa Leboa – English online dictionary published by TshwaneDJe HLT. This presentation discusses the advantages of each dictionary and shows that they all can empower their users but that none of the three dictionaries can cater for everybody in all situations because there is no such thing as THE dictionary that provides a solution for everything.Let's detail the opportunities offered by the online dictionary market in three areas:

- Search Engine Optimization (SEO): why dictionary content is a marvellous resource to answer a wide range of queries in search tools such as Google, Bing, Yandex or Baidu,
- Reaching local markets worldwide with bilingual content,
- User Generated Content: an unmissable resource.

> **Dictionaries and their influence on language purification in minority languages. The case of Frisian**

KUIP, FRITS VAN DER

10 – Lexicography of Lesser Used or Non-State Languages

In literature, scepticism on the effect of language propaganda is dominant. Researchers observe that it is almost impossible to stop lexical interferences from becoming current in standard languages such as Dutch (or Southern Dutch in Belgium) through language purification literature or through language-related articles and transmissions in the media or (last but not least) through concise dictionaries.

The question we have to ask ourselves is, whether the dictionaries' influence in a minority language such as Frisian is limited as well. Most speakers of Frisian are, as far as writing is concerned, illiterate in their

own language. They are not accustomed to written Frisian word forms and unsure when it comes to how their language should be written correctly. A Frisian speaker will be more inclined to consult a dictionary when writing something in his own language, than a speaker of a majority language or a national language would do. On the basis of that assumption, you would expect that including purisms and avoiding or marking interferences in dictionaries, would significantly affect the written language at least.
In this survey, I looked at four loan-words, including the loan-translations and purisms (if any) that go with them. I compared the occurrences (and non-occurrences) of these words as dictionary entries to their respective frequencies of occurrences in two major databases.
On the one hand, we see that, throughout the years, the purisms included in the dictionaries perform considerably better than the equivalent loan-words and loan-translations. The purisms not in the dictionaries perform considerably worse. On the other hand, we notice a trend among writers of Frisian to use interference words in the last few decennia. So, at first glance, dictionaries seem to have influenced language purification. However, one cannot tell for how long that will be the case. It will depend on speakers' attitudes towards their language. After all, it is very difficult to control a language as has been proven in the case of Dutch, and the same might hold for Frisian.

> Mobile phone dictionaries for small languages: the Whitesands electronic dictionary

McELVENNY, JAMES

10 – Lexicography of Lesser Used or Non-State Languages

This poster presentation reports on work to develop an electronic dictionary of Whitesands (Austronesian; Tanna Island, Vanuatu) which can be stored on and accessed through mobile phones. In the presentation we will outline some of the benefits of mobile phone electronic dictionaries for speakers of small languages, look at some of the difficulties that we had to overcome in preparing the dictionary, and discuss the reception of the dictionary in the Whitesands community.

> Scottish Lexicography: Major Resources in Minority Languages

PIKE, LORNA AND ROBINSON, CHRISTINE

10 – Lexicography of Lesser Used or Non-State Languages

This paper focuses on current aspects of the lexicography of two minority languages in Scotland, Scottish Gaelic and Scots, and looks at two projects at either end of the lexicographical spectrum: Faclair na Gàidhlig, an on-

line full historical dictionary of Gaelic and the new edition of the *Concise Scots Dictionary* (*CSD*, 1985), a one-volume derived dictionary of Scots.
A brief outline of the history of both languages is given. Each in turn was the dominant language in Scotland until both were replaced by English.
The paper looks at how Scotland's minority languages have benefited from the skills of the Scots who contributed to English lexicography. Sir James Murray, first Editor of the *Oxford English Dictionary* (*OED*), pioneered the application of historical principles to English lexicography and his colleague, Sir William Craigie, applied those same principles to the *Dictionary of the Older Scottish Tongue* (*DOST*) which covers the Scots language from the 12th century to 1700. These skills are now being transferred into Scottish Gaelic.
Faclair na Gàidhlig will be an on-line historical dictionary of Gaelic compiled on similar principles to *OED* and *DOST*. The major challenge in establishing a project of this magnitude is to create a lexicographical tradition as effectively and efficiently as possible. The paper outlines the approach adopted. A draft noun entry is examined with discussion of entry structure and organisation.
Scots is equipped with two historical dictionaries, *DOST* and its modern counterpart the *Scottish National Dictionary* (*SND*). *CSD* is a one-volume distillation of these works. The second edition will use a more user-friendly structure and update coverage to the 21st century. Sample entries are examined.
Scottish lexicography will continue to build on its historical tradition providing Gaelic and Scots with resources comparable to English.

> **The Lexicographic Work of Euskaltzaindia – The Basque Language Academy 1984-2009**

SAGARNA, ANDONI

10 – Lexicography of Lesser Used or Non-State Languages

The Academy started developing a standard variety of Basque Language in 1968. In 1983 the Academy created a commission of lexicography, and in 1984 approved a long-term plan for the development of dictionaries. That plan included the following projects:

1 The General Basque Dictionary (GBD), which should be a compilation of the lexicon used in the publications until 1970.
2 A lexicology project, whose aim was to study the formation of words in Basque.
3 A compilation of the lexicon used in current publications

The corpus of GBD contains 6 million text words, The first result of the project was a dictionary of 16 volumes in paper format. Since October

of 2009 this dictionary is available online at the address http://www.euskaltzaindia.net/oeh . The result of the compilation of the lexicon used in current publications is The Statistical Corpus of the Twentieth Century that contains 4,658,036 words from 6,351 pieces of text. It is available online at the address http://www.euskaracorpusa.net/XXmendea/Konts_arrunta_fr.html.

In 1992 the Academy created a commission to prepare The Unified Dictionary of the Basque Language. The General Basque Dictionary and The Statistical Corpus of the Twentieth Century are precisely the information sources that provide insight into the use of words. In 2000, was published a list of standardized 20,000 words and in 2008 a second edition collected a total of 29,000 words. By the end of 2011 the list should contain about 40,000 forms. Regardless of the publications on paper, the unified dictionary is available at http://www.euskaltzaindia.net/hiztegibatua

> **Lexicography of a Non-State Language: The Case of Burgenland Romani**

SCHRAMMEL, BARBARA AND RADER, ASTRID

10 – Lexicography of Lesser Used or Non-State Languages

Burgenland Romani (henceforth BR) is spoken in Burgenland, the easternmost province of Austria. Until recently BR was an exclusively oral language. However, active language use of BR has almost totally ceased in the second half of the 20th century. The self-organisation of the group from the 1990s onwards led to a new appreciation of the language, which is now accepted as the primary identity marker. This new interest in their own language and culture entails the desire for the revival, maintenance and spread of BR. One aspect of language planning in BR concerns the functional expansion of the language into acrolectal domains where it has never been used before.

BR is lexicographically documented in two different media, i.e. in ROMLEX (henceforth RL), which is an extendible multi-dialectal lexical database with a freely accessible web-interface (http://romani.uni-graz.at/romlex/) and a print dictionary. RL is intended as a tool for comprehensive lexical documentation of BR. At the same time, it is a practical, low-threshold tool for text producers. The print dictionary, on the other hand, primarily serves an emblematic purpose. Given the differing purposes of RL and the print dictionary, different strategies are used in lexicographic decision-making. Roughly speaking, RL favours an inclusive descriptive approach while the print dictionary is rather restrictive and follows normative principles. The paper discusses decisions taken with respect to orthography, lemma selection and meaning for RL and the print

dictionary, respectively. We are highlighting lexicographic phenomena, such as increased polysemy, generic usage of terms and heavy borrowing, which are typical of the functional expansion process of stateless minority languages.

> An Overall View about Lexicography Production for the Friulian Language

TOFFOLI, DONATO

10 – Lexicography of Lesser Used or Non-State Languages

Researches about distinctive characteristics of Friulan language began in the second half of the Eighteenth Century. There was a need to have lexicography instruments that could help to understand and convert the lexicon heritage of the language actually spoken in large parts of Friuli, Friulian language, in the quite unknown linguistic code of the new born Italian state, the Italian language.
Some of these lexicographic works still exist and are nowadays very important: for example '*Il Nuovo Pirona*', a formidable tool for the dialectological work; also other interesting tools were printed such as, for example, '*Vocabolario della lingua friulana*' by Giorgio Faggin, or '*Dizionario pratico illustrato italiano-friulano*' by Maria Tore Barbina.
A meaningful change happened in the 90's when the first law to safeguard and promote the Friulian language was approved and the first body for the linguistic policy, l'*Osservatorio Regionale della Lingua e Culture Friulane* (OLF), was founded. The lexicographic work began to be more structured.
A great work has been done for computer medium: a spelling corrector: '*Coretôr ortografic furlan*' by '*Informazione Friulana*' Cooperative; a dictionary on CD '*Dizionari Ortografic Furlan*' (DOF) by Alessandro Carrozzo.
Lately The most important lexicographic work is '*Grant Dizionari Bilengâl Talian Furlan*' (GBDTF), by *Consorzio Friûl Lenghe* 2000, based on a prestigious Italian model such as the '*Grande Dizionario dell'Uso della Lingua Italiana*', by professor Tullio de Mauro that coordinated the Friulian version too.

> The Klagenfurt lexicon database for sign languages as a web application: LedaSila, a free sign language database for international use

KRAMMER, KLAUDIA

11 – Sign Language

The Klagenfurt online database for sign languages 'LedaSila' (Lexical Database for Sign Languages, http://ledasila.uni-klu.ac.at/) is designed in such a way that it is possible to present all the information which can be found in any good monolingual or bilingual (printed) dictionary. It offers

the possibility to enter semantic, pragmatic as well as morphosyntactic information. Furthermore, a detailed analysis of the non-manual and manual parameters of a sign is possible. LedaSila offers the possibility to search for any information already contained in it (including single signs or formational parameters), to document a sign language, or analyse it linguistically. The search function is accessible to all internet users. When using the database for sign language documentation and/or analyses, an authorisation from the Centre for Sign Language and Deaf Communication in Klagenfurt is required.

When using LedaSila for documentation and/or analysis of a sign, a user does not have to follow a specific order when entering the data. Furthermore, the user is free to decide whether to enter data only in one field (e.g. semantics or region) or to do a full analysis of the sign. A special feature of LedaSila is the possibility to add new categories and values at any time. This is especially important for an analysis tool which is designed to be used internationally. This feature ensures that all categories and values needed for a specific sign language are available.

LedaSila can be used free of charge for non-commercial deaf and scientific issues. The database is hosted on a server of the University of Klagenfurt. All information (including videos) is stored directly on the web server. This means that using LedaSila comes with zero administration. The international sign language linguistic community is invited to use this easily manageable database.

> The Danish Sign Language Dictionary

KRISTOFFERSEN, JETTE H. AND TROELSGÅRD, THOMAS

11 – Sign Language

The entries of the The Danish Sign Language Dictionary have four sections:

Entry header: In this section the sign headword is shown as a photo and a gloss. The first occurring location and handshape of the sign are shown as icons.

Video window: By default the base form of the sign headword is shown. Other types of videos are rendered in this window, but activated by clicking play buttons in different sections of the entry.

Meanings window: In this section the meanings of the sign are shown. The meaning description includes: Danish equivalents, a description of the sign's usage (for function signs) and mouth movement, cross-references to synonyms etc., information about restricted use, and example sentences. Semantically opaque compounds with the sign are shown below the regular meanings.

Additional information: This section holds cross-references to homonyms and to common frozen forms of the sign (only for classifier entries). In addition to this, frequent co-occurrences with the sign are shown in this section.

The signs in the The Danish Sign Language Dictionary can be looked up through:

Handshape: Particular handshapes for the active and the passive hand can be specified. There are 65 searchable handshapes.

Location: Location is chosen from a page with 15 location icons, representing locations on or near the body.

Text: Text searches are performed both on Danish equivalents, sign glosses and example sentences (both transcriptions and translations). This enables users to find signs that are not themselves lemmas in the dictionary, but appear in example sentences.

Topic: Topics can be chosen as search criteria from a list of 70 topics.

> The first national Dutch Sign Language (NGT) Dictionary in book form: Van Dale Basiswoordenboek Nederlandse Gebarentaal

SCHERMER, TRUDE AND KOOLHOF, CORLINE

11 – Sign Language

In October 2009 the first national Dutch sign language (NGT) dictionary in book form was published by Van Dale publishers. (Schermer,Koolhof (eds) 2009). The content of the book is produced by the national centre for NGT and for sign language lexicography, the Dutch Sign Centre, and is based on 25 years of research into the lexicon of Dutch Sign Language (NGT) which we will describe briefly in our paper (Schermer 1990, 2004). Subsequently we will describe organisation and content of the Van Dale dictionary which contains 3000 standard signs with illustrations ordered alphabetically by using a glos as lemma.. In addition to the Van Dale dictionary in book form an online NGT dictionary is available on our website (Schermer,Koolhof,Muller 2010) which offers both search features: alphabetically and via handshape/location. Each entry in the Van Dale dictionary contains further information: an example sentence of how the sign is used and grammatical information about the non manual features and type of verb. We will show examples from both dictionaries, discussing the dilemma's we faced and the solutions we opted for in the making of this dictionary.