The *European Network of e-Lexicography* (ENeL)

Tanneke Schoonheim

On October 11ᵗʰ 2013, the kick-off meeting of the *European Network of e-Lexicography* (ENeL) project took place in Brussels. This meeting was the outcome of an idea ventilated a year and a half earlier, in March 2012 in Berlin, at the *European Workshop on Future Standards in Lexicography*. The workshop participants then confirmed the imperative to coordinate and harmonise research in the field of (electronic) lexicography across Europe, namely to share expertise relating to standards, discuss new methodologies in lexicography that fully exploit the possibilities of the digital medium, reflect on the pan-European nature of the languages of Europe and attain a wider audience.

A proposal was written by a team of researchers from the Instituut voor Nederlandse Lexicologie, the Fryske Akademy and the Berlin-Brandenburgische Akademie für Wissenschaften, partners from all over Europe were found and in May 2013 the action was approved of by COST (European Cooperation in Science and Technology), which is an EU framework supporting cooperation among scientists and researchers across Europe. This means that up to the end of 2017 public funds are available for meetings, workshops, training schools and conferences with regard to the European Network of e-Lexicography.

This presentation of ENeL explains the background, aim, structure and impact of the network, and its meetings, workshops, training schools and scientific missions.

General background
The background for setting up the network was that computers and the internet have seriously changed the conditions for the production and reception of dictionaries. The internet offers entirely new possibilities for developing and presenting dictionary information, such as with the integration of sound, maps or video, and various novel ways of interacting with dictionary users. For editors of scholarly dictionaries the new medium is not only a source of inspiration, it also generates new and serious challenges that demand cooperation and standardization on various levels:

a. Through the internet scholarly dictionaries can potentially reach large audiences. However, at present scholarly dictionaries providing reliable information are often not easy to find and are hard to decode for a non-academic audience; 'traditional' dictionary users tend to use easily accessible non-academic and user-generated dictionaries. In order to let a larger audience benefit from higher quality dictionary content, we should bridge the gap between the general public and scholarly dictionaries by improving access to these dictionaries and making them more widely known.

b. Most scholarly dictionary projects make their products available on the internet or have plans for going online. All of them find themselves confronted with similar problems relating to technologies for producing lexicographical content, presentation, interaction with users, etc. Most dictionaries take different approaches to and find different solutions for these problems. So far European and international cooperation in these fields has been restricted to bilateral collaborations. There is a clear need for a broader and more systematic exchange of expertise and for the establishment of common standards and solutions.

c. In the past years, innovative forms of electronic dictionaries

Editor | **Ilan Kernerman**

KDICTIONARIES

**Joint Symposia of
JACET Society of English
Lexicography and
Kansai English
Lexicography Circle
30 May 2014
Aichi University, Nagoya**

The joint symposia of the JACET Society of English Lexicography and the Kensai English Lexicography Society (KELC) was organized by Professors Michihisa Tsukamoto and Yoshihito Kamakura from Aichi University in Nagoya.

If Japan is counted among lexicographic powerhouses, these organizations are two of the driving forces behind it, contributing to the improvement and promotion of theoretical and practical lexicography.

JACET Society of English Lexicography is a group under the auspices of the Japan Association of College English Teachers. It was co-founded in 1995 by Professors Minoru Murata and Kosei Minamide, and chaired by Professor Murata from 1995 to 2006 and Professor Kaoru Akasu from 2007 to 2013. The society holds an annual workshop (including a joint one with its sister society JACET Vocabulary Acquisition Research Group once in three years), and has played a key role in organizing the two Asialex conferences held in Japan (Tokyo 2003 and Kyoto 2011). KELC was founded in 2003 with a view to read literature ▶

have appeared that no longer resemble traditional paper dictionaries but try to fully exploit the new possibilities of the digital medium. Though serious attempts have already been made at embedding electronic lexicography into a theoretical framework, an up-to-date research paradigm and common standards for electronic lexicography are still lacking, as are common standards and cooperation for interlinking content of digitized dictionaries and innovative e-dictionaries.

d. The digital medium offers the potential for a new type of lexicography that no longer views languages as isolated entities. Language migration has always been part of human history, yet this is often not properly reflected in dictionaries. Many dictionaries have their origins in the 19th century, where a national perspective on language prevailed. Consequently, the information in the dictionaries is mostly presented from the point of view of a single language, not sufficiently taking into account similar or related developments in other European languages. This view of languages often dominates also modern dictionaries. There is a clear need for a common approach to e-lexicography that forms the basis for a new type of lexicography that fully embraces the pan-European nature of much of the vocabularies of languages spoken in Europe.

Aim of the network

The aim of the European Network of e-Lexicography is to increase, co-ordinate and harmonise research in the field of e-lexicography and to make authoritative information on the languages of Europe easily accessible.

The network will:

- Make lexical knowledge of small and large languages available in a European dictionary portal. This portal will serve as the central reference point for all dictionary users who look for reliable, authoritative dictionary information on the languages of Europe and their histories on the internet.
- Enable cooperation and the exchange of resources, technologies and experience in e-lexicography and provide support for dictionaries that are not yet online.
- Discuss and aim at establishing standards for innovative e-dictionaries that fully exploit the possibilities of the digital medium.
- Establish new ways of representing the common heritage of European languages by developing shared editorial practices and by interconnecting already existing information.

Working groups

Four working groups are the main vehicles of delivering the scientific program. Each consists of lexicographers and computational linguists from different countries and includes both experienced researchers and young ones. The working groups are led by a chair and vice-chair, who also provide annual reports on its activities.

All researchers affiliated to an institution in a COST country can join the network. There is also room for researchers from countries neighbouring a COST country and from Argentina, New Zealand and South Africa.

Working group 1.
Integrated interface for European dictionary content

Working group 1 investigates how authoritative dictionary information on the languages of Europe can be made accessible to both the general and academic public. Chair of the working group is Anne Dykstra (Netherlands) and vice-chair is Bob Boelhouwer (Netherlands).

The working group will:

- Set up a European dictionary portal, which will give information on scholarly dictionaries of the languages of Europe and provide access to these dictionaries. Different parameters will be considered, e.g. dictionary type, language covered, digitally-born versus retro-digitized.
- Investigate the possibilities of interlinking the contents of European dictionaries.
- Investigate user requirements with respect to the presentation of dictionary content.
- Investigate the possible involvement of users in the creation of dictionary content.

Working group 2.
Retro-digitized dictionaries

In working group 2 the focus is on the digitization of paper dictionaries. It intends to set up guidelines and standards for turning paper dictionaries into digital format. Chair of the working group is Vera Hildenbrandt (Germany) and vice-chair is Vladimír Benko (Slovakia).

Work will be carried out on:

- The development of standards for encoding of information and the description of relevant categories for print dictionaries.
- Presenting an overview of software for the conversion of physical layout information to logical information.
- The investigation of relevant information categories to be added to the dictionary in order to make its content more readily accessible and interoperable.
- The development of a work plan for

digitisation, including parameters necessary for estimating costs.
- The investigation of possible use of dictionary content for computational linguistic applications.
- The organisation of a training school on standard tools and methods for retro-digitising dictionaries in 2015.

Working group 3.
Innovative e-dictionaries

The scope of working group 3 is the development of digitally-born dictionaries, focusing in particular on the latest developments in e-lexicography and the interface between lexicography and computational linguistics. Chair of the working group is Simon Krek (Slovenia) and vice-chair is Carole Tiberius (Netherlands). Work will be carried out on:
- The description of the workflow for corpus-based lexicography.
- Providing an overview of existing software needed to set up this workflow.
- The use of dictionary writing systems.
- The analysis of the possible impact of automatic acquisition of lexical data.
- The analysis of the interface between dictionary and computational lexica (cf. wordnets) and syntactically and semantically annotated corpora (cf. FrameNet, SemCor, Senseval).
- The investigation of the possible use of dictionary content for language technology applications.
- The organisation of a training school on innovative approaches in e-lexicography in 2016.

Working group 4.
Lexicography and lexicology from a pan-European perspective

Working group 4 investigates how the pan-European nature of the vocabularies of the languages of Europe can be represented in single-language dictionaries and within the European dictionary portal. This is particularly relevant to studying the multiple dimensions of borrowing, i.e. the migration and re-migration of words and meanings across the languages of Europe. Chair of the working group is Eveline Wandl-Vogt (Austria) and vice-chair is Phil Withington (UK).
This working group will:
- Develop ways in which already existing information from single language dictionaries can be displayed and interlinked to represent more adequately their common European heritage.
- Develop editorial guidelines for the integration of European information into more traditional dictionaries as well as innovative ones.

- Find new applications for the very large amount of interconnected dictionary information from the European dictionary portal in the field of digital humanities.
- The organisation of a training school in lexicography and lexicology from a pan-European perspective in 2017.

Training schools and short term scientific missions

One of the responsibilities of the network is to organise training schools and to provide so-called short-time scientific missions (STSMs). The three training schools mentioned above will be organised by members of the relevant working groups, in close cooperation with the training school manager of the network, Rute Costa (Portugal). Participants of the training schools will get their travel and accommodation expenses reimbursed by the network.

The first training school will be held in 2015 on the subject of retro-digitization. More information on this training school will be available at the time of the network meeting in July 2014 and will also be published on the website: http://www.elexicography.eu/events/training-schools/.

As part of an STSM researchers visit established dictionary projects and centres of excellence in a country other than their own. The visit period may take from five days to three months and can range from discussions and demonstrations to deeper involvement in the activities of the centres visited, in accordance with the needs of the researchers. Travel and accommodation expenses of approved missions are reimbursed by the network. There is room for at least four STSMs in each year, which brings the total of STSMs to take place within the duration of the project (till October 2017) to sixteen or more.

STSMs are especially meant to allow young researchers to build up their own networks, and they will facilitate and increase the capacity for research in the field of e-lexicography. Tanneke Schoonheim (Netherlands) is appointed as the manager of these missions. For more information on the missions and how to apply, see http://www.elexicography.eu/events/workshops/.

Impact of the network

The network will allow the exchange of knowledge and expertise in the field of e-lexicography. It is open to all scholarly dictionaries in Europe irrespective of their previous experience with e-lexicography. The network will facilitate the coordination and progressive expansion and standardization of research activity through the work of the four working groups and

▶
about lexicography, lexical semantics and corpus linguistics, and deepen the understanding of recent research findings in these fields of study. Coordinated by Professor Kosei Minamide, the circle meets bi-monthly to discuss the issues and insights raised in selected books, and its members have translated and brought three of these to publication in Japan: *Phraseology* (Cowie, ed. 2001), *Words and Phrases* (Michael Stubbs, 2001), and *Lexicography: An Introduction* (Howard Jackson, 2002). The regular meeting of the JACET group is in Tokyo and that of KELC is in Kyoto. Located between the two cities, Nagoya was a fine venue for the joint symposia. The meeting began with a special session on 'Challenges Facing English Dictionary Editors', with the participation of the chief editors of four major English-Japanese learners' dictionaries, and continued with sessions on 'EFL Dictionaries: Description and Use' and 'Collocation and English-Japanese Dictionaries'. Altogether there were a dozen talks, attended by 80 participants. The symposium included a book display and was followed by a buffet-style party at the university cafeteria.

Shigeru Yamada
Chair, JACET Society of English Lexicography

through close cooperation with the Euralex and e-Lex conferences.

The sustainability of the network will be promoted through the European dictionary portal and through the publication of reports and roadmaps concerning future standards for e-lexicography on the network's website. This website is intended to become the first source of consultation by lexicographers planning to create new e-dictionaries.

In the European dictionary portal reliable information on the languages of Europe will be made easily available to everyone. The portal will therefore be a trustworthy alternative to the many user-generated dictionaries on the internet. In addition, whereas traditional printed dictionaries are often inaccessible to people with a visual impairment (set in very small type, heavily abbreviated for space-saving reasons) the portal will give visually impaired people easy access to these dictionaries with screen readers.

The development of common standards will save time and money. New e-dictionaries or dictionaries that are due to go online will no longer have to develop their own standards but will be able to refer to the publications of the network. In particular dictionaries of small languages will benefit from such common and easily-available standards, as they usually do not have the means to develop their own approaches. The network can therefore substantially contribute to regional or minority languages.

The overall quality of the dictionaries will be improved. The new editorial methods and practises to be developed will reflect more realistically the language situation in Europe and the historical development and interaction of European languages, in particular the migration of words and meanings across Europe. Lexicographers and lexicological researchers, especially young ones, will increase their knowledge and skills by participating in working groups, training schools and conferences. Yvonne Luther (Germany) is appointed as manager for young researchers and/or female researchers to make sure that they get the right chances in the network.

The large amount of data connected in the European dictionary portal will enable new lines of research in the field of digital humanities that could have not been carried out on the basis of isolated language resources, e.g. the spread of technological innovation by studying the appearance of relevant words in the vocabularies of the languages in Europe.

Structure of the network

The network is driven by a management committee, comprising an elected chair and vice-chair and up to two representatives from each COST country. The chair is Martin Everaert (Netherlands) and vice-chair is Iztok Kosem (Slovenia). The management committee meets every six months and is responsible for co-ordinating the activities of the working groups, budget planning and the allocation of funds, organising training schools and conferences. It will monitor progress in relation to the scientific focus and work plan in relation to the achievement of milestones.

The steering group is responsible for preparing annual reports on the work of the network, overseeing the development and maintenance of the network website, communicating with the COST office and monitoring its procedures. It comprises the chair and vice-chair of the management committee, the chair and vice-chair of each working group, and the managers for the training schools, STSMs and young/female researchers. The steering group meets every three months, having started in January 2014 in Leiden, followed in April in Vienna, in July in Bolzano and in October in a place yet unknown.

At present the network comprises representatives from 29 COST countries: Austria, Belgium, Bulgaria, Croatia, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Latvia, Netherlands, Norway, Poland, Portugal, Romania, Serbia, Slovak Republic, Slovenia, Spain, Sweden, Switzerland and the United Kingdom.

Members of the network meet every six months, preferably in combination with other conferences regarding lexicography or computational linguistics, such as Euralex, e-Lex and LREC. On these occasions there will always be meetings of the working groups, the management committee and the steering group, followed by a plenary session for all members on the progress made. The next network meeting is on July 19-20 in Bolzano (Italy), directly following the Euralex conference. The first meeting of 2015 is due in Vienna (Austria) in January or February and the second meeting may be connected to e-Lex 2015 in July in Herstmonceux Castle (UK).

For more information on ENeL, please visit our website: www.elexicography.eu.

EUROPEAN NETWORK OF **e-LEXICOGRAPHY**

Multilingual Linked Open Data for Enterprises MLODE 2014

The second workshop on Multilingual Linked Open Data for Enterprises (MLODE 2014) will be held as part of the SEMANTiCS conference in Leipzig on 2 September 2014. It will bring together developers, data producers, academia and enterprises from various fields of linguistics, natural language processing (NLP) and information technology to present and discuss principles, case studies and best practices for representing, publishing and linking linguistic data collections, including corpora, dictionaries, lexical networks, translation memories, thesauri, etc.

As Semantic Web research progresses, interest by practitioners, industry and infrastructure providers operating across language barriers increases. The Linked Open Data (LOD) community is enthusiastic about the possibilities offered by new, vast multilingual resources. While it is clear that the Semantic Web is not a panacea, it has matured into a technology capable of addressing specific real world problems of globalization faced by the industry and the governmental sector.

The first MLODE workshop (2012) was successful in establishing a highly-productive interdisciplinary network that enables researchers to share experiences on how to sustainably manage and interlink huge amounts of language data through Semantic Web technologies, providing scientific resources that enhance the development of more precise models and applications for research on Linked Data. A major output of MLODE 2012 was the improvement of the technical viability of the Linguistic Linked Open Data (LLOD) cloud (http://sabre2012.infai.org/mlode/).

The goal of MLODE 2014 is to compare technologies and datasets developed before and in parallel to Linked Data, RDF and OWL, and to offer industrial participants an overview of the technologies ready for exploitation. The workshop will focus on the following topics:

- **Unifying the Dictionary**
  The continuation and centralization of various efforts to compile a multilingual dictionary integrating heterogeneous sources. Open questions on focusing the research and crowd-sourcing a multilingual core lexicon will be tackled with actual corporate use cases, combining the know-how of the assembled research communities and industry professionals associated with the SEMANTiCS conference. In this context, the DBpedia project is eager to contribute to the creation, hosting and maintenance of final resources, guaranteeing persistence, impact and visibility of the outcome.
- **Building LLOD-aware NLP services**
  The development of resources and services, including comprehensive structured metadata, with the final goal of granting effortless programmatic access to multilingual natural language processing services based on open data.
- **Generating Linked Data for Language Resources**
  The continuation of the initiative began at MLODE 2012 to convert, aggregate and publish Language Resources and extend the LLOD-cloud, which has grown significantly, contributing linguistic data to the NLP service infrastructure (http://linguistics.okfn.org/resources/llod/).

MLODE 2014 also aims to present cases for, and barriers against, industry participation in Linked Data for NLP and content internationalization and localization, discuss best practices on how to channel feedback from industry to open-source and academic communities, and produce a roadmap for Linked Data & Language Technology in Europe.

http://mlode2014.blogs.aksw.org/

Sebastian Hellmann, Bettina Klimek
Universität Leipzig

MLODE organization
LIDER (http://lider-project.eu)
Linked Data for Language Technology W3C Community Group (http://www.w3.org/community/ld4lt/)
Working Group for Open Data in Linguistics (http://linguistics.okfn.org)
K Dictionaries (http://kdictionaries.com)
Ontotext (http://ontotext.com)
Clarin (http://clarin.eu)
Language Science Press (http://langsci-press.org/)
DBpedia Association (http://dbpedia.org/Association)
InfAI (http://infai.org)
NLP2RDF (http://nlp2rdf.org)

KDNews 2014

INTERNS @ KD
Zulema Badanes Canet | Eva Prats Balaguer | Miguel Angel Bordas – Universitat Jaume I, Spain
Bettina Klimek – Universität Leipzig, Germany
Dılara Özlem Güneren – Université de Lille 3, France
Kseniya Egorova | Anna Bespyatova – Saint-Petersburg State Economic University, Russia
Alex Milton de Carvalho | Bruna Rafael Neira Munoz – UNESP, Brazil
Louis Albrecht – Université de Lorraine (ATILF CNRS), France

KD in CHINA
The Commercial Press International Company (Beijing) will collaborate with K Dictionaries on dictionary publication in China. The cooperation will be launched with a new edition of *Random House Kernerman Webster's College Dictionary* and include a Chinese bilingual pocket dictionary series, starting with French and Japanese.
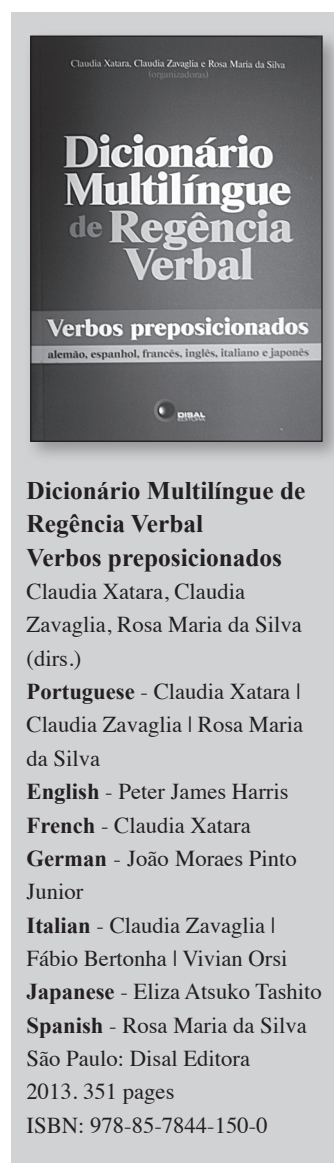
CONFERENCE SPONSOR
LREC – Reykjavik, Iceland
EURALEX – Bolzano, Italy
MLODE – Leipzig, Germany

Claudia Xatara, Claudia Zavaglia, Rosa Maria da Silva (dirs.).
Dicionário Multilíngue de Regência Verbal – Verbos preposicionados

It is a well-known fact that collocations and prepositional regency are the hardest part of learning a foreign language, since the arbitrariness of the linguistic sign also affects the combinability of the signs; besides, theoretical phraseology expanded a long time ago its object of study beyond set phrases and proverbs, extending its interest to light verb constructions, routine formulae, interjections, insults, curses, etc. The compilation of dictionaries has not always duly accompanied this progress, and, in many languages, good dictionaries of collocations are still missing, or, when available, appear much later, in comparison to other types of dictionaries. The fact that the boundaries between what belongs to phraseology and what belongs to syntax are not unanimously accepted is perhaps one of the reasons why, in many languages, not enough attention has been granted until recently to the production of dictionaries of propositional schemes or verbal valences.

Although verbal valency is a syntactic phenomenon, it depends largely on the meaning of each verb, since it cannot be predicted only from formal rules. Such rules may explain the mechanics of "arguments" like subject and object for transitive verbs, but when prepositional complements appear the casuistry explodes. A verb may have figurative meanings that alter completely the valency, e.g. Portuguese *dar* (to give), which is trivalent *par excellence*, loses its arguments in some sequences that are neither ditransitive nor idiomatic set phrases. In Spanish, *la terraza da al mar* (*the terrace gives to the sea *the terrace overlooks the sea*) has no direct objet; or *dio con la cabeza contra el muro* (*he gave with the head against the wall *he beat his head against the wall*) has neither indirect nor direct object. These examples are not necessarily set-phrases, although they may undergo a metaphor that integrates them into an idiom, like in Spanish *dar en el clavo* (*to give in the nail *to hit the nail*). Therefore, the figurative meanings are linked to a prepositional regency as variable and as whimsical as English phrasal verbs, with a certain degree of phraseological fixation that affects not only the seemingly arbitrary union between a given verb and a given preposition, but also the verbal valences, which are no longer the same as in the "literal" verb.

The team led by Brazilian phraseologists

Claudia Xatara, Claudia Zavaglia and Rosa Maria da Silva made this *Dicionário Multilíngue de Regência Verbal – Verbos preposicionadas* facing the difficult task of making an inventory of these events, ordering and describing them, comparing each construction in seven languages (Portuguese, English, French, German, Italian, Japanese, Spanish), ordering them alphabetically from the Portuguese version. Starting from 6,000 verbs taken from Borba's grammatical dictionary (Francisco Da Silva Borba, *Dicionário grammatical de verbos do português contemporâneo*. São Paulo: Editora UNESP, 1991), and taking information from large general dictionaries, like *Aurélio* (Aurélio Buarque de Holanda Ferreira, *Dicionário Aurélio da Língua Portuguesa*. Curitiba: Editora Positivo, 1975, 1999, 2010) or *Houaiss* (*Grande Dicionário Houaiss da Língua Portuguesa*. Rio de Janeiro: Instituto Antônio Houaiss, 2001, 2003), data were checked and compared with the *Corpus Textual Electrônico do Laboratório de Lexicografia* of UNESP (São Paulo State University), gathering 1,200 Brazilian Portuguese words, clarifying and describing their construction regency. The authors' starting assumption is that prepositional complements are also mandatory arguments of the verb in certain constructions (e.g. *simpatizar com* *sympathize with, *rezar por* *pray for), which are called in Portuguese *verbos preposicionados* (*prepositional verbs). An explanatory paraphrase and a Portuguese definition of each entry are provided, with examples of real usage, as well as equivalents in the six other languages (with the good idea of adding the transliteration into Latin alphabet for Japanese hiragana and katakana characters).

One of the most representative entries is *dar* (give), which has 35 sub-entries, where the regular valences of the literal meanings of the verb inevitably meet together with the not so literal ones, and with several more or less idiomatic combinations, such as *dar com alguem* (*give with someone *to meet someone*), equivalent to Spanish and German reflexive constructions requiring another verb (*encontrarse con alguien*; *mit sich jemandem treffen* *to find oneself with someone*), and to a completely different metaphor in Italian (*imbattersi in qualcuno* *bump into someone*).

Since a foreign language learner cannot know beforehand the boundaries between

categories which are still controversial even for specialists, it is clear that a prototypical user of a bilingual dictionary needs a reference work where all this kind of information is available simultaneously, without knowing in advance whether the searched sequences are syntactic, lexical or phraseological. Since most of the valences involve prepositional complements, and we cannot expect the learner to distinguish *a priori* between verbal valency and prepositional regency, the dictionary must enable the user to access both phenomena as they occur naturally in speech: "mixed together".

The Spanish verb *consentir* (*to allow) is trivalent (A **consiente** B *a* C) (*someone consents something to someone), but it is also possible to have (A **consiente en** B), with no addressee, a construction that is possible also in French but without changing the preposition (*consentir à qqch.*) while, in Italian, the ellipsis of the addressee requires to change the verb and the preposition at the same time (**acconsentire a**: **permettere** *a qualcuno* **di**).

In this sense, verbal valences also belong to the field of lexical combinations, and need a detailed lexicographical treatment, including the lexical and prepositional environment of each verb, not only for their literal meaning, but also for their figurative and idiomatic values. The cross-linguistic dimension means the predictable valency in one language may have, in another language, an equivalent whose argument is a prepositional complement instead of a direct object. Besides, each preposition can govern another morpho-syntactic case, thus the possibilities are multiplied.

For example, let's have a look at DREAM + NAME OF ACTION: in Italian the dreamed action is represented by a direct object (*sogna viaggiare* *he dreams travelling*), whereas in Spanish and Portuguese it is a prepositional complement of company (*soñar con viajar / sonhar com viajar* *to dream with travelling*) and in French and German there is a genitive construction (*rêver de… Traumen von* *to dream of*).

Another clear example of this apparent arbitrariness is "FALLING IN LOVE", which in Spanish requires *de* (genitive construction: *enamorarse de*), while English requires *with* (committative construction: *to fall in love with*), German requires *in* (locative construction: *sich verlieben in*), and Portuguese requires *por* (ablative construction: *se apaixonar por*). In this sense, Spanish *andar detrás de N* (*to walk after N *to be looking for (something)*) allows us to suspect that the difference between collocations, idioms and phrasal verbs is not as clear and objective as is often thought.

The entries are presented as follows:

DAR (31) POR algo/alguém (*perceber a ausência de*)
→ *quando derem **por** mim será tarde demais*
A: *jemandem etwas vermissen* **0**
E: **echar en falta** *algo/alguien*
F: *remarquer l'absence **de***
In: *note the absence **of***
It.: *accorgersi **di***
J: **ga iru (inai) no ni ki ga tsuku**

DAR (34) algo POR algo (*desfazer-se; vender*)
→ *só darei o meu Picasso **por** uma fortuna equivalente á sua beleza*
A: *etwas **für** etwas verkaufen*
E: *vender **por***
F: *donner **contre**; donner **pour***
In: *to give **up for***
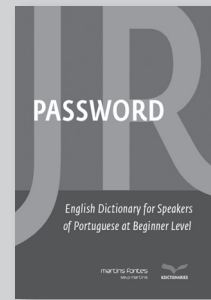It.: *dare **per**; vendere **per***
J: **wo** *to korihiki suru*

The applications to natural language processing and/or machine translation are also important. For example, the Google Translate tool (http://translate.google.com/, accessed May 2014), though based on translation memory and statistics, shows a dramatically wrong result for *dar por alguém* (**by giving someone*) or for Spanish *echar en falta a* (**hacks to take*) instead of *to notice the absence of*. Results are no better with the *Word Reference* tool (http://wordreference.com/, accessed May 2014). The same can be said about the great majority of such constructions, except when, incidentally, English coincides literally with the Portuguese (or Spanish, etc) form.

This dictionary is, thus, an excellent tool, not only for foreign learners (and teachers) of the Portuguese language, but also for research purposes in contrastive linguistics in a field that has been a kind of 'no man's land' for too many years: the borderline between syntax, idioms and lexicology.

Antonio Pamies Bertrán
Dept. Lingüística General y Teoría de la Literatura, Universidad de Granada
antonio.pamies@gmail.com

Reverso Context: Redefining dictionaries and language tools

Théo Hoffenberg

Théo Hoffenberg is CEO of Softissimo Inc. He is an engineer by training (graduate of École Polytechnique in Paris), an entrepreneur by profession (founder of Reverso among others), a designer and architect of NLP tools and content by passion, and an amateur linguist.
theo@reverso.com

The last issue of this publication had a brief article by Ilan Kernerman entitled Dictionary *n*. *Obsolete*? (KDN21, 2013). Although the title is certainly provocative, it's quite clear that today's dictionaries are not exactly what they used to be. In the same issue Colin McIntosh wrote about the new definition for 'book' in *Cambridge Advanced Learner's Dictionary*, which focused more on the content than the physical form we used to associate it with (ibid.).

Combining the two approaches, we could say the dictionary as we knew it (i.e. a book consisting of a list of entries indexed in alphabetical order, containing definitions or translations, examples of usage, compounds, etc) is probably already obsolete. Nevertheless, we are likely to still use the word 'dictionary' to refer to a very different concept, just as we still use 'telephone' for something that no longer resembles the large contraption with a rotary dial it represented 30 or 40 years ago.

Soon the word 'dictionary' will likely refer to a tool that helps us find the most appropriate choice in a certain context, offering users easy access to meanings and translations of words and phrases, along with relevant examples of usage, etc.

Within this framework, Reverso presents a new approach to translation aids that in a few years might be understood as a dictionary, but for the time being can have different names: a new type of example-based dictionary; a bilingual concordancer; a search engine for large bilingual texts (*bitexts* in NLP jargon) aligned at word and phrase level; a bilingual aligner providing translation for relatively frequent sequences of words; a provider of frequent wording suggestions and their translations; an analyzing tool that applies linguistic principles to big data; a terminology checker based on balanced corpora.

These descriptions may seem intimidating at first, and may even bring to mind a Rube Goldberg machine, or a white elephant. However, in practice, novices and experts alike find this approach efficient and easy to use. Therefore we call it simply Reverso Context (RC). While the text below offers some insight on how our idea of a 'dictionary' works, readers are invited to also experience it firsthand in order to fully appreciate the innovative features of this linguistic tool.

Changes

Dictionaries of previous generations had various limitations. Firstly, the total number of characters must fit an acceptable volume (say below 2,000 pages for the large ones). Thus, reducing the content size and eliminating redundancy and inconsistencies soon became a huge task that required enormous work from authors and editors looking to fit comprehensive data within restricted space. Secondly, the print was in black and white, occasionally including only one additional color.

The use of such dictionaries demanded the reader's active participation to interpret signs or symbols (e.g. ~ to replace the headword, -> for cross-reference), as well as to cross-reference the indicators suggested (subject, object type, preposition use, etc) with the actual context in order to choose the most appropriate meaning. Cross-checking in the opposite direction was also in the hands of the user. Moreover, looking up phrases such as *not at all*, *je m'en vais* or *pas tout de suite* often proved to be a difficult task.

At that time, there were no search engines, and intensive users of foreign languages were fewer (though perhaps more motivated). Nowadays, many people are used to search engines and machine translation, hence some laziness or higher expectations on their part. When searching for appropriate vocabulary, modern users expect answers to be instant, precise and varied.

Users are also increasingly used to ask questions in whichever way they come to mind, without rephrasing or adapting to query syntax, and still be able to obtain relevant answers.

In 2000, based on comprehensive dictionaries from Collins, we made a big step by putting computer power into use to enhance the user experience for dictionary look-up: no more ~ to replace the headword or other such abbreviations; use of color to identify components of an entry (blue for source language, black for target language, green for domain indicators, red for grammar, etc); direct access to compounds; and full-text search to find examples in both directions (for example, *faire miroiter* appears as the translation for *dangle* even though *dangle* is not given

as a translation for *faire miroiter*). This feature was initially implemented in our dedicated software environment called *Lexibase*, which included Collins bilingual and monolingual dictionaries. The software has been updated and is still in use today, available for Internet, intranet and PC, and the same environment has been applied to many more dictionaries since then.

Despite its extensiveness, the content itself was originally designed with the intention of producing a book. This means that the variety of the examples and the coverage of derivations, among other components, were limited, and focus was put more on avoiding redundancy rather than on expanding coverage.

Dictionaries of this type were not only limited, but also extremely costly to develop, because of the 100% human factor and the strict editorial rules and compactness. As a result, most dictionaries for non-major language pairs, such as French-Arabic or French-Japanese, never reached the comprehensiveness of those for English-Spanish, for example.

But even the largest dictionary content originally designed for a print edition cannot provide the full coverage expected today in terms of examples, derivatives or context, let alone up-to-date vocabulary and technical terms.

Examples

Let's take some concrete examples of the benefits of this advanced language tool that allows users to communicate in languages that are not their mother tongue.

Suppose a French-speaking person wants to translate *je m'en vais* into English. Wouldn't it be nice to type this text in an entry box, and get translation suggestions including examples that use both the searched item and the suggested results in context? What if the same possibilities existed for a Spanish speaker looking to translate *me voy*? This is precisely what RC is about.

Users these days are accustomed to getting relevant results in a blink of an eye and effortlessly. In this sense, RC caters to "pampered" users that no longer wish to lemmatize such phrases. After all, knowing that *je m'en vais* stems from *s'en aller* isn't obvious, and searching through the sub-entries of *aller* for the verb's pronominal or reflexive forms can be tiresome. Additionally, if you happen to be a linguist, you know that important information is often lost through lemmatization, as not all verb forms relate to the original meaning of the root.

This same example can be observed from another viewpoint. Although native speakers' intuition allows them to know that *je m'en vais* can have very different meanings according to context, new learners or even proficient non-native speakers may find it difficult to grasp the different nuances of this phrase. In fact, the tone of this expression can range from neutral to aggressive and threatening, and its meaning varies when it precedes a verb, in which case it expresses a will to take action.

Looking up *me voy* in the large Collins English-Spanish dictionary, for example, one might automatically switch to full-text search to display relevant examples containing this text, but still not find the direct translation of the phrase itself. In addition, RC offers more than 8,000 short texts containing the item, of which over 1,000 are aligned to *I'm going* and 400 to *I'm leaving* and *I'm off*.

RC is also particularly useful for finding examples of usage and translation of phrases that cannot be translated independently. Take for example the phrase *shy of* + number or quantity, which can be translated as *un peu moins de*, meaning that a certain amount is less than expected. The examples enable users to find the most suitable expression for each particular situation. The same applies for other words, such as *sorted*, *get sorted*, *get things sorted*, *get myself sorted*, etc.

For the linguist, RC offers more interesting features, allowing to identify trends or validate theories and lexicons. It responds to questions such as: What are the most frequent translations for this word or phrase? Which frequently used phrases contain this word? What does this word translate into when not in this phrase?

Taking an example, a quick look-up of *upside* in RC will show that most examples of usage are related to *upside down*. A more advanced search provides translations excluding the phrase *upside down*. Then, if a certain phrase (e.g. *upside risks*) is too widely represented, it can be excluded from the search. Alternatively, simply looking up *an upside* will provide translations of *upside* as a noun.

When translating *siège* from French to English, words such as *seat*, *siege* and *headquarters* may come to mind. Although one may think that *seat* is the most generic translation and that *headquarters* is used mainly as part of the compound *siège social*, a search with RC would show that *headquarters* is by far the most commonly used equivalent. Moreover, the "-" option can be applied in the RC (*siège -{siège social}*) to check if *siège* is translated into *headquarters* even when it is not part of *siège social*.

Non-natives who are proficient in a foreign language often need assistance

to validate their choice of words. This process of producing a coherent translation is what linguists describe as *encoding*, for *production* purposes.

One way to do it is to use online dictionaries, starting with bilingual dictionaries from the source language into the target language and then searching the definition or synonyms in the target language to find the most appropriate one in context. Another option would be to look up the definition or synonym of the assumed translation and see whether it is appropriate, also using the reverse translation (translating back into the source language).

For example, to translate *acharné* a French-English dictionary would provide equivalents such as *fierce*, *bitter*, *relentless* and *unremitting* as first proposals, with only a few examples. To search further, one could look for synonyms for *relentless* and find *ruthless*, *unrelenting* and *uncompromising*. However, in order to find the best translation, one should have a near-native level of English, or at least read the definitions or the "back translations".

Moreover, if one were to use this adjective as part of *travail acharné*, results could be surprising. The RC search shows that it is widely used, and that its translation is *hard work*, although *hard* is not among the proposed translations or synonyms for *acharné*.

If the *relentless* translation is chosen, one could check sentences containing the first one in the source language, and the second one in the target language, finding more than ten relevant examples.

Conclusion

Bilingual concordancers have already proven to answer certain requirements that dictionaries could not meet. This explains why their use has spread in recent years and why RC could profit from being easy-to-handle by average users though providing more advanced features. RC will continue to innovate and push towards the dictionary of the future. First, by improving and diversifying content resources, adding new and varied corpora that encompass diverse fields and language levels, and expanding coverage to both written and spoken language. Second, by introducing new ways to customize the user's search experience. For example, large organizations that have voluminous corpora are already able to prioritize their content with more pertinent features, and subject domain, regional variants and the language level will also be possible to filter in the future. Last but not least, RC will strive to maintain high quality when dealing with large data volumes thanks to automatic cleaning scripts as well as processing user feedback. With this, we hope to be to the dictionary what a smartphone is to the old telephone today.

An introduction to iFinger and Clarify Language Service

Knut Haga

iFinger is a provider of digital dictionaries integrated in Microsoft Windows environment. The first version of the iFinger software was released in 2000, with the main goal of offering convenient look-up solutions in high-quality dictionaries. The portfolio varies from glossaries to unabridged monolingual and bilingual dictionaries from HarperCollins, Merriam-Webster, K Dictionaries, Pons and Cappelan Damm, as well as iFinger's own terminology for the medical, technical and legal domains.

Since its inception, over 3.5 million users have accessed this dictionary software from CNET's Download.com. It enables tailored application for multiple users in the corporate, educational and governmental sectors. At present iFinger has more than 200,000 users in the educational sector in Norway and more than 50,000 users in corporate and governmental sector.

Overall, the demand for language services is growing constantly. iFinger aims to meet this demand by developing new cloud-based language tools that will be available as native solutions for all common operating systems. This service is branded as Clarify and is offered for free to the general public, with premium content available through annual subscription to corporate and government users.

The initial launch in June 2014 offers a free dictionary service for 30 languages, including 670 dictionaries covering the languages of 3.4 billon people. There are mobile apps for iOS, to be followed soon by apps for Android and Windows Mobile, as well as by new services for Machine Translation and Text To Speech.

This entire activity is being transferred from iFinger to the new Clarify service, which is designed for unlimited global growth.

http://clarifylanguage.com
http://ifinger.com

ASIALEX 2015, Hong Kong

The 9th International Conference of ASIALEX will be hosted from 25 to 27 June 2015 by the Hong Kong Polytechnic University.

ASIALEX, the Asian Association of Lexicography, is approaching its eighteenth anniversary. Founded in the Dictionaries in Asia conference at Hong Kong University of Science and Technology in 1997, the association has made great efforts to bring linguists, lexicographers and dictionary users together for the development of lexicography in Asia. It has acted as a catalyst for scholarly cooperation and information-sharing in a wide range of areas: dictionary compilation, dictionary critique, bilingual lexicography, learners' dictionaries, user study, Asian language study and computational lexicography. The biennial conference is a major event in the development of ASIALEX. The table below records the landmarks.

Following the vein, the theme of ASIALEX 2015 is *Words, Corpora and Dictionaries: Innovations in reference science*, with the following topics:

- dictionary and ELT
- dictionary compilation in a digital era
- the role of corpora in reference science
- dictionary, corpus and Asian languages
- multimedia and multifunction of the dictionary
- dictionary and culture
- terminology, phraseology and neologisms

The papers presented at ASIALEX 2015 will be published in the conference proceedings, and selected papers will be recommended to the association's refereed journal *Lexicography: Journal of Asialex*.

Important dates:
- Abstract submission date: 15 February 2015
- Notification of paper acceptance: 30 March 2015
- Deadline for early registration: 30 April 2015
- Deadline for paper submission: 10 May 2015
- Conference dates: 25-27 June 2015
Submission guidelines:
- The medium of the conference is English
- The abstract should not exceed 300 words (excluding references)
- Abstract submission is online at http://asialex2015.engl.polyu.edu.hk/

It is our great honour and pleasure to bring ASIALEX back to its birthplace, Hong Kong, in 2015. Great changes have taken place in lexicography, in the city, in the region and around the world in the last 18 years. We welcome lexicography experts, linguistic educationists, young scholars and dictionary enthusiasts to join us in this regional and international event to witness the growth and changes in lexicography in Asia.

See you in Hong Kong in 2015!
http://asialex2015.engl.polyu.edu.hk/

Dr Lan Li, Convener
Department of English
Hong Kong Polytechnic University
lan.li@polyu.edu.hk

Conference	Year	Venue	Theme	President
inauguration	1997	Hong Kong	*Dictionaries in Asia: Research and pedagogical implications*	
1st	1999	Guangzhou, China	*National experience in lexicography or dictionary compilation and bilingual lexicography*	Huang Jianhua (China)
2nd	2001	Seoul, Korea	*Asian bilingualism and the dictionary*	Sangsup Lee (Korea)
3rd	2003	Tokyo, Japan	*Dictionaries and language learning: How can dictionaries help human and machine learning?*	Minoru Murata (Japan)
4th	2005	Singapore	*Words in Asian cultural contexts*	Anne Pakir (Singapore)
5th	2007	Chennai, India	*Asian lexicography: Retrospect and prospect*	V. Jayadevan (India)
6th	2009	Bangkok, Thailand	*Dictionary in education*	Jirapa Vitayapirak (Thailand)
7th	2011	Kyoto, Japan	*Lexicography: Theoretical and practical perspectives*	Zafar Iqbal (Pakistan)
8th	2013	Bali, Indonesia	*Lexicography and dictionaries in the information age*	Yukio Tono (Japan)
9th	2015	Hong Kong	*Words, corpora and dictionaries: Innovations in reference science*	Chu-Ren Huang (Hong Kong)

Opening Statement
- Lexicography in Asia: Its future and challenges | Yukio Tono

Papers
- Sketch Engine: Ten years on | Adam Kilgarriff
- Lexical Markup Framework | Gil Francopoulo and Chu-Ren Huang
- Korean and English 'dictionary' questions: What does the public want to know? | Susanne Bae and Hilary Nesi
- "Bottom-up" approach in making verb entries in a monolingual Indonesian learner's dictionary | Dora Amalia
- Towards improved coverage of Southeast Asian Englishes in the *Oxford English Dictionary* | Danica Salazar

Research Report
- An analysis of the smartphone dictionary app market | Christoper Winstock and Young-kuk Jeong

Book Reviews
- *The Chinese-English Encarta Dictionary* (2011) | Lan Li
- *The Bloomsbury Companion to Lexicography* (2013) | Vincent Ooi

Welcome to our new website

soft launch throughout 2014

http://**kdictionaries**.com