Kernerman
DICTIONARY News
Number 24 • July 2016
kdictionaries.com/kdn

Towards Globalex

The GLOBALEX Workshop on Lexicographic Resources and Human Language Technology (http://ailab.ijs.si/globalex/) took place as part of LREC 2016 at Portorož, Slovenia on May 24 and constituted the first live step in forming an overall global constellation for lexicography. The initiative was launched nine months earlier at a meeting held during the fourth eLex conference in the UK in August 2015, and has drawn the support of lexicographic associations worldwide.

The full-day workshop was sponsored by the associations for lexicography of Africa, Asia, Australasia, Europe and North America (Afrilex, Asialex, Australex, Euralex, DSNA), and the eLex conference series on electronic lexicography in the 21st century. It set out to explore standards for lexicographic resources and their incorporation in new language technology and other solutions as part of knowledge systems and collaborative intelligence. The workshop was attended by about 60 participants, included 16 twenty-minute sessions and concluded with a roundtable about the future of Globalex.

globaLex

The core idea of Globalex is to work on lexicography in global contexts and bring together different segments that operate on their own – on regional, topical or any other level – to cooperate.

It is hoped that Globalex can facilitate knowledge sharing and cooperation among its members and with others concerned with language and language technology, promote the creation, research, exchange, dissemination, integration and usage of lexicographic resources and solutions, and enhance interoperability with the academia and industry worldwide.

The roundtable featured short interventions by a representative of each organization, including one by video and another by skype, presenting their association and vision of Globalex, followed by a discussion with the audience. The main issues concerned the aims and obstacles facing Globalex, its organization, operation and meetings. The conference models ranged from dedicating a section to Globalex at the continental conferences, and alternating Globalex conferences with those of the different associations, to holding Globalex conferences on their own every few years.

The organizers have agreed to contribute to the new Globalex website http://globalex.link/, which begins operation this month. More details appear on page 4, and a reprint of Towards Peoplex, from 1997, is available on page 18.

Ilan Kernerman

Editor | **Ilan Kernerman**

KDICTIONARIES

K DICTIONARIES LTD
8 Nahum Hanavi Street
Tel Aviv 6350310 Israel
+972-3-5468102
kdl@kdictionaries.com
http://kdictionaries.com

ISSN 1565-4745

The African association for Lexicography (Afrilex) was established in 1995 after a feasibility study for a lexicographical institute for Southern Africa indicated a keen interest in a unifying body among lexicographers and members of related professions. Dr Reinhard R.K. Hartmann chaired the inaugural meeting, and officially announced the birth of a new member of the Lex family.

Afrilex is managed by a Board elected biennially by the members present at a General Meeting of the association. Membership is open to individuals and institutions who have an interest in lexicography. The current membership stands at 60 individuals and 8 corporate members. The board consists of the president, vice-president, secretary, treasurer, four non-officers and the conference convener.

The aims of Afrilex include the promotion and coordination of research, study and teaching of lexicography by means of publishing a journal, *Lexikos*, and other appropriate literature, organizing regular conferences and seminars that offer opportunities for exchange of ideas and for mutual stimulus to researchers and practitioners in the field of lexicography, and facilitating the participation in tutorials and training courses.

Afrilex seeks to develop cooperation with other international associations for lexicography as well as with local associations that are interested in the study of language.

The 21st annual International Conference of Afrilex is held in July 2016 in Tzaneen, South Africa.

Lexikos (ISSN 2224-0039) is the official mouthpiece of Afrilex, the editor being an ex-officio member of the Board. All contributions are indexed by the Thomson Reuters Web of Science Citation Index and are freely available online (http://lexikos.journals.ac.za/pub/).

In its first twenty years of existence Afrilex has bestowed Honorary Membership on the following members: Prof. A.C. Nkabinde, Prof. Rufus Gouws, Dr Johan du Plessis, and Dr Mariëtta Alberts.

http://afrilex.africanlanguages.com/homelex.html/

AFRILEX

The Asian Association for Lexicography (Asialex) was established at the initiative of Gregory James and Amy Chi on 29 March 1997, during the Dictionaries in Asia conference at Hong Kong University of Science and Technology, with the aim of fostering scholarly and professional activities in the field of lexicography and facilitating the exchange of information and ideas through meetings, publications, etc. Membership is open to any person or institution.

The first executive board was elected at that inauguration meeting, and the President, HUANG Jianhua, convened the first conference in Guangzhou (1999). From then on, elections were not held again, and usually the convener of each conference was named president for two years, until the voting process was renewed in Kyoto 2011.

Asialex is governed by an executive committee that is elected for two-year terms, consisting of a president, vice-president, secretary, treasurer, and three more members as well as four ex-officio members including the immediate past president, journal editor, and conveners of next two conferences.

Lexicography – Journal of Asialex is published biannually since 2014 by Springer, in print and online, and membership is connected to the journal subscription. Until then, the activity of Asialex focused almost entirely on holding biennial international conferences. In addition to conference proceedings, a newsletter appeared in the first years and collections of papers from two conferences were published as well. Since 2015, conferences started to be held once a year, with the tenth taking place in Manila 2016, and the next one due in Guangzhou in 2017.

The challenges facing Asialex and achieving its goals are inherent in Asia's non-homogeneity on multiple levels. This vast geographical region is composed of different areas often disconnected from each other, and its enormous linguistic diversity is often under-resourced, under-researched or under-represented. Traditionally Asialex has had stronger presence of the eastern parts and much less of central, south and western Asia. Overcoming the challenges would uncover and leverage their resourcefullness.

http://asialex.org/

ASIALEX
The Asian Association for Lexicography

The Australasian Association for Lexicography (Australex) was founded in 1990 as a companion association to Euralex. It is committed to the development of lexicography in all languages of the Australasian region. Its interests include:

- dictionaries of all kinds
- the theory of lexicography
- the history of lexicography
- the practice of dictionary-making
- dictionary use
- endangered languages
- Revivalistics
- terminology and terminography
- corpus lexicography
- computational lexicography
- sign language
- lexicology

Membership consists mainly of people from Australia, New Zealand and the Pacific Islands, but also from many other countries, including Japan, South Africa, Spain, the UK and Zambia. Australex includes career lexicographers, students of lexicography, researchers into dictionaries, publishers, teachers and people who just like dictionaries.

The association is governed by a committee of 10 members, who are elected every two years during the biennial conference. It consists of a President, Vice-President, Secretary, Treasurer, five officers and the immediate past President. Membership is free.

Until 2009, meetings were held regularly every one or two years, in addition to specific conferences (e.g. on Australian placenames of indigenous origins) and workshops (e.g. on dictionary writing). Since then conferences have been held biennially, in either Australia or New Zealand. The next conference is planned for August 2017 in the Cook Islands. It is hoped that this location will extend the range of Australex and involve speakers of more language groups, particularly endangered ones. The conferences are usually small, which has the benefit of promoting close collaboration and networking, with the opportunity for delegates to attend most of the presentations. One or more student bursaries are offered to help with conference attendance.

Australex has one self-publication of peer-reviewed papers from its 2013 conference, entitled *Endangered Words and Signs of Revival* (2014).

http://adelaide.edu.au/australex/

AUSTRALEX

The Dictionary Society of North America (DSNA) was founded in 1975 to foster scholarly and professional activities relating to dictionaries, lexicography, and lexicology and to bring together people interested in the making, study, collection, and use of dictionaries. DSNA's principal activities include a biennial conference, a biannual newsletter, a website, and a journal. DSNA sponsors a lexicography course at the Linguistic Society of America Summer Institute and funds a fellowship for a student to attend. Occasional informal local meetings for members have begun, and outreach efforts to promote better public understanding of lexicography are underway. DSNA is a member of the American Council of Learned Societies.

A president, vice-president, and executive secretary are DSNA's officers and with four elected at-large members constitute the executive board, with the immediate past president an ex-officio member. The journal and newsletter editors regularly participate in the conference calls of the board and report to DSNA's publications committee each month. Other committees address finance, nominations, membership, etc. Currently, DSNA enrolls about 250 individual and institutional members.

Dictionaries—DSNA's journal—aims to represent the best research in lexicography and lexicology, including history, theory, and practice of lexicography, and the design and use of dictionaries and related works of reference. It publishes peer-reviewed articles, invited contributions, book reviews, reports of reference works in progress, and occasional forums. Published annually, it has in recent years averaged 285 pages; a move to biannual publication is under consideration. The journal is indexed in MLA Bibliography, Linguistics and Language Behavior Abstracts, and Linguistics Abstracts; all issues are accessible through Project MUSE.

DSNA derives its revenue from membership fees, journal royalties, and gifts. Student memberships are free of charge. Both financially and programmatically the biennial conferences are the responsibility of the host institution.

http://dictionarysociety.com/

The series of conferences on electronic lexicography in the 21st century (eLex) was started in 2009 by Sylviane Granger in response to this emerging field. Initially, the conference (at Louvain-la-Neuve, Belgium) was conceived as a one-off event, however its success and calls from the lexicographic community for a follow-up prompted Iztok Kosem and Simon Krek to turn it into a biennial conference series. The subsequent conferences in Bled, Slovenia (2011), Tallinn, Estonia (2013), and Herstmonceux Castle, UK (2015) thus focused on different topical issues and attracted increasing numbers of participants from all over the world.

As eLex is not an association, it does not have an official board, a membership fee, etc, but there is an unofficial committee consisting of chairs of organisational committees of previous conferences. The committee offers local organisers of the next eLex conference advice on and help with organisational matters. Furthermore, members of the committee maintain the eLex website, which provides links to the webpages of all previous conferences, including proceedings, programmes and other relevant information on related activities.

The eLex conferences have always promoted interdisciplinarity, bringing together specialists in dictionary publishing, corpus lexicography, software development, language technology, language learning and teaching, translation studies, and theoretical and applied linguistics. There has also been a constant effort put into the dissemination of topical developments and issues in (electronic) lexicography among members of the community worldwide. An important part of achieving this goal have been videorecordings of the presentations and round tables which have been made freely available on the conference websites.

The next eLex conference will be hosted by the Institute of the Dutch Language and held in Leiden, the Netherlands, in the second half of September 2017. Further announcements with more detailed information will be made on the eLex website and posted on relevant mailing lists.

https://elex.link/

The European Assoction for Lexicography (Euralex) brings together people working in lexicography and related fields. In the rapidly-changing world of language analysis and language description, it provides a forum for the exchange of relevant ideas. Though based in Europe, Euralex has a worldwide reach and a worldwide membership. Its members include lexicographers, reference publishers, corpus linguists, computational linguists, academics working in relevant disciplines, software developers, and anyone with a lively interest in language.

Euralex holds a major conference every two years, and also sponsors smaller events on specific areas within the broader field. The first conference was held in Exeter, UK, in 1983 and since then there have been conferences on a regular basis in 13 different countries all over Europe – the 17th to be held in Tbilisi, Georgia, in September this year. Euralex has created a digitized version of all the papers from its past conferences, freely available from its website.

Euralex maintains a discussion list for the exchange of views on anything of interest to people working in lexicography and related fields. The list is public and not limited to members. It also maintains a public Facebook page.

In cooperation with Oxford University Press, Euralex is responsible for the *International Journal of Lexicography*, a leading peer-reviewed academic journal that appears four times a year. Interdisciplinary as well as international, it is concerned with all aspects of lexicography, including issues of design, compilation and use, and with dictionaries of all languages, though the chief focus is on dictionaries of the major European languages – monolingual and bilingual, synchronic and diachronic, pedagogical and encyclopedic.

Euralex is governed by an executive board consisting of up to nine elected members, including four principal officers (President, Vice President, Secretary-Treasurer and Assistant Secretary-Treasurer), elected at each general meeting from among its members. The general meeting is held in connection with the biennial conference.

http://euralex.org/

GLOBALEX 2016 workshop summary and next steps

GLOBALEX 2016 organizing committee

Andrea Abel
(Euralex, Vice-President)

Ilan Kerenerman
(Asialex, President; workshop co-chair)

Steve Kleinedler
(DSNA, Vice-President)

Iztok Kosem
(eLex, Co-Organizer)

Simon Krek
(eLex, Co-Organizer; workshop co-chair)

Julia Miller
(Australex, President)

Maropeng Victor Mojela
(Afrilex, President)

Danie. J. Prinsloo
(Afrilex, Vice-President)

Rachel Edita O. Roxas
(Asialex, Vice-President)

Lars Trap-Jensen
(Euralex, President)

Luanne von Schneidemesser
(DSNA, President)

Michael Walsh
(Australex, Vice-President)

GLOBALEX preparatory board 2016-2018

Edward Finegan, Ilan Kernerman, Iztok Kosem, Simon Krek, Julia Miller, Maropeng Victor Mojela, Lars Trap-Jensen

http://globalex.link/

The GLOBALEX 2016 Workshop on Lexicographic Resources for Human Language Technology was held as part of the Language Resources and Evaluation Conference (LREC 2016) at Portorož, Slovenia, on 24 May 2016, with approximately 60 participants, and constituted the first step in forming a global network for lexicography. The organizing committee consisted of the presidents and vice-presidents of Afrilex, Asialex, Australex, DSNA and Euralex, and two co-organizers of eLex, with the actual working group comprising Ilan Kernerman, Iztok Kosem, Lars Trap-Jensen and Simon Krek. The program committee (including about 40 members) selected 16 papers out of 24 submissions, each having 15 minutes for presentation plus 5 minutes for questions. More details are available on the workshop website http://ailab.ijs.si/globalex/, and the proceedings are at http://lrec-conf.org/proceedings/lrec2016/workshops/LREC2016Workshop-GLOBALEX_Proceedings.pdf/.

The workshop closed with a roundtable discussion about next steps for GLOBALEX. It lasted about 90 minutes and had about 30 participants. The first part was moderated by Simon and consisted of a brief presentation of each member organization, including four in person, by Sonja Bosch (Afrilex), Ilan (Asialex), Lars (Euralex) and Iztok (eLex), a video by Julia Miller (Australex), and live Skype participation by Edward Finegan (DSNA). The second part was moderated by Ilan in discussion with those present, including Ed by Skype. The following main points emerged:

- All the associations and the individuals present welcomed the formation of GLOBALEX as an umbrella constellation to enhance worldwide cooperation and exchange on lexicography
- The organizations began to cooperate in this vein by co-organizing this GLOBALEX workshop
- The five continental associations have reached a consensus on jointly running a new GLOBALEX website and sharing its hosting costs
- The new GLOBALEX website http://globalex.link/ will go live in mid-2016
- The site will function as a repository and will link conference proceedings, presentations, slides and videos, post news, announcements, etc.
- It is hoped GLOBALEX can operate in a lean non-bureaucratic fashion; still a steering committee is needed, including a representative of each body
- Participation will be open to any local, special topic, or other lexicography-minded community, and will serve to promote the members' interests and activities
- Holding conferences could be handled by combining models of various activities, such as:
  - Video-record talks at different conferences and post them on the GLOBALEX website
  - Minimize or avoid meetings scheduled across the world at the same time
  - Promote meeting in person whenever and wherever possible
  - Have GLOBALEX sessions as part of other conferences
  - Welcome the interest expressed in principle by Afrilex and Australex to hold GLOBALEX sessions as part of their 2017 conferences
  - Advance the DSNA offer to hold a GLOBALEX session at its 2017 meeting
  - Organize virtual conferences with the aim to keep costs low, facilitate the participation of those from distant places or with fewer resources, take advantage of technology, and attain wider dissemination of information worldwide
  - Ultimately hold face-to-face GLOBALEX *Olympic* conferences every few years

Work has already begun to set the ground for implementing such ideas and planning further. For example, talk is underway to leverage the coinciding DSNA and Asialex conferences in June 2017 for cross-broadcasting between dinner in Barbados and morning plenary in Guangzhou. In June, the organizing committee's four working-group members were joined by one liaison each from Afrilex (Victor Mojela), Australex (Julia Miller), and DSNA (Ed) to continue operating together as a preparatory board for the formation of GLOBALEX. Besides the immediate task of getting the new website up and running, the main work will focus on setting a framework for GLOBALEX to function in full cooperation with all its members and to their benefit on a worldwide scale, deciding what shape such body might have, defining its strategic policy, decision-making process, etc. It is encouraging that the initiative launched last year has already had such an overwhelmingly welcoming response and is starting to develop.

XVII Euralex International Congress

The XVII Euralex International Congress of the European Association for Lexicography will be held on 6-10 September 2016 in Tbilisi by the Lexicographic Center of Ivane Javakhishvili Tbilisi State University.

With the theme of *Lexicography and Linguistic Diversity*, the main objective of the congress is to highlight the importance of lexicography for the preservation of linguistic diversity and the promotion of cultural and scientific ties among different cultures and nations. Other objectives are to emphasize the role of lexicography as a rapidly developing interdisciplinary branch of science – incorporating multiple components, *viz*. semantic theories, corpus-based methods, techniques for natural language processing, e-lexicography, etc – and to explore the current status of lexicography as merely a *craft* or rather a full-fledged scholarly discipline destined to fulfill multiple important missions in our rapidly developing multicultural and multilingual world. In addition, the Tbilisi congress aims at further popularization and sustainable development of lexicography in Georgia.

The Programme Committee has selected 115 papers by 190 authors from around the world. The papers were anonymously reviewed by at least two members of the Scientific Committee. Keynote lectures will be delivered by Jost Gippert, Patrick Hanks, Robert Ilson, Pius ten Hacken and Geoffrey Williams, and a round-table discussion will be moderated by Thierry Fontenelle. The programme includes also parallel sessions, software demonstrations, pre-congress tutorials and specialized workshops, a book and software exhibition, and social events.
http://Euralex2016.tsu.ge/

Tinatin Margalitadze is director of the Lexicographic Centre at Tbilisi State University and convener of the Euralex conference.
http://margaliti.com/index_en.htm/

The Lexicographic Centre at Tbilisi State University

In May 2011 the Academic Council and the Council of Representatives of Ivane Javakhishvili Tbilisi State University took the decision to grant the Lexicographic Centre at the university the status of University Centre for Bilingual Lexicography. The decision was part of the process of consolidating the role of Georgian as the national language and the language of science in Georgia. This is particularly important at this current moment in the history of Georgia as, following the restoration of independence in 1991, the Georgian language regained its function as the national language and began to develop and adapt to the realities of contemporary life, incorporating words and expressions connected with international politics and diplomacy, market economy and judicial procedures, as well as military, scientific and technical terms.

The Lexicographic Centre (LC) was originally established as an independent entity by the Department of English Philology back in 1995 and included the editorial team of the *Comprehensive English-Georgian Dictionary* (CEGD) that has been in place since the 1980's. The aim was to edit and prepare CEGD for publication in volumes, and 14 volumes have been published so far, covering a total of 2,380 pages. In 2009 the LC started to work on an electronic platform for CEGD and in 2010 the *Comprehensive English-Georgian Online Dictionary* was posted on the Internet (http://margaliti.ge/eng/index.htm/). The online version is based on the published volumes and includes 110,000 entries.

In 2008 it was transformed into a faculty-level centre within the Faculty of Humanities and started the compilation of a series of specialized dictionaries. *English-Georgian Online Military Dictionary* (http://mil.dict.ge/) was created in 2009 at the request of the Ministry of Defence. Then, the LC editors compiled *English-Georgian Online Biology Dictionary* in 2011-2013 (http://bio.dict.ge/) and *English-Georgian Online Dictionary of Technical Terminology* in 2014-2016 (http://tech.dict.ge/), both funded by Shota Rustaveli National Science Foundation.

One of the LC goals is the promotion of bilingual lexicography of Georgian and European languages, for which purpose MA and PhD programs were launched. In 2011 a joint MA program was set up with the Department of the Italian Language and the work on a *New Italian-Georgian Learner's Dictionary* is the first one underway. The LC plans to initiate bilingual projects for other European languages, including old ones such as Gothic and Old English. In 2012 the LC started to work on a new project, Parallel Corpus of English-Georgian Scientific Texts (http://corp.dict.ge/).

The LC pays great attention to the promotion of lexicography as a branch of science. With that end in view, it delivers public lectures, gives presentations, has trainings with teachers of foreign languages, arranges contests, and aims to provide adequate education in this field. The LC has been a key force in transforming the approach of the authorities towards lexicography in Georgia. It was one of the initiators of setting up a State Committee for the Enhancement of Lexicography in Georgia at the Ministry of Education and Science. The Committee is developing the National Programme of Lexicography, which is intended to compile Georgian explanatory, historical and specialist terminological dictionaries, and to promote bilingual and electronic lexicography.

Asialex 2017 in Guangzhou

XU Hai
Convener of Asialex 2017
Center for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies

The 11th International Conference of The Asian Association for Lexicography (ASIALEX) will be held at Guangdong University of Foreign Studies (GDUFS) in Guangzhou, China on June 10-12, 2017. This conference will mark the 20th anniversary of ASIALEX. Being the host of the First International Conference of ASIALEX, we are very pleased to bring it back to this location for this landmark event after it has traveled around nine Asian countries and regions over the past twenty years.

The theme of ASIALEX 2017 is *Lexicography in Asia: Challenges, Innovations and Prospects*. The main topics are as follows:

- electronic and digital revolution in lexicography
- computer corpus lexicography
- bilingual lexicography
- pedagogical lexicography
- metalexicography
- dictionary use studies
- dictionary and culture
- dictionary as discourse
- phraseology
- neologisms
- terminology

To respond to the challenges of the corpus revolution and the digital revolution in lexicography, lexicographers, linguists, language professionals and publishers from across Asia and worldwide need to work together to share information, knowledge and experience, and to encourage innovation in lexicographic studies and practice. Our conference aims to provide such a platform.

GDUFS is a major internationalized university known for its global-minded faculty members and students and its research on language, literature, culture, trade and strategic studies. With 21 foreign languages available, it is the only university in South China to offer such a great variety of programs, and its foreign language and literature courses as academic disciplines are among the finest nationwide. It boasts a key national research center for humanities and social sciences, Center for Linguistics and Applied Linguistics, which is under the auspices of the Ministry of Education and conducts leading research in lexicography and applied linguistics. The members in the center have published with the top presses in China including Commercial Press and Foreign Language Teaching and Research Press (FLTRP), and with leading scholarly journals including the *International Journal of Lexicography*, *Lexikos* and *Lexicography – Journal of ASIALEX*.

We are very pleased that HUANG Jianhua, the first President of ASIALEX who convened that first conference in GDUFS in 1999, will be one of the plenary speakers in ASIALEX 2017. Prof Huang is a renowned lexicographer and has recently completed a 16-year gigantic dictionary project – *Grand Dictionnaire Chinois-Français Contemporain* (FLTRP, Beijing, 2014) – the largest Chinese-French dictionary in the world. The other keynote speakers include Andrea Abel, of EURAC and currently Vice-President of Euralex; Julia Miller, of the University of Adelaide and President of Australex; and Michael Rundell, of Lexicography Masterclass and Macmillan Dictionaries.

We hope that you will join us in celebrating this 20th anniversary of ASIALEX and look forward to welcoming you in Guangzhou next June!

Nineteenth-Century Lexicography Conference, 2018

A conference on 19th-century lexicography – Between Science and Fiction – will be held at Stanford University on 6-7 April 2018, with an aim to explore the following issues:

How can we understand the making of monolingual and multilingual dictionaries in the 19th century? Were lexicographers in conversation with philologists, seeing their work as science and to be undertaken collaboratively, by teams of scientific observers? Or were they utopian thinkers, trying to create new languages or to form writers and speakers who would use old languages in new ways? How are the prescriptive and the descriptive intertwined in their work? What evidence do dictionaries in different languages offer to answer these questions? What were lexicographers' personal motives for their work? What role, if any, did nationalistic enterprises play in the planning and execution of these texts? What were the historical factors – as regards technology or thought – that led to the flourishing of lexicography in this period? And what brings this phenomenon to scholars' attention now?

Please send 300-word abstracts to Sarah Ogilvie (sogilvie@stanford.edu) and Gabriella Safran (gsafran@stanford.edu) by 1 September 2016.

Chinalex and lexicographic activities in China

Yihua Zhang

Abstract
The China Association for Lexicography (Chinalex) plays an important role in Chinese lexicography. This article offers a general introduction to Chinalex and sets forth the functions it has performed in the lexicographic activities and characteristics of lexicographic pactice in China, followed by a presentation of a new generation of learners' dictionaries and attempts made in computer-aided lexicography.
Keywords: China Association for Lexicography, Chinalex, lexicographic activities, learner's dictionary, computer-aided lexicography

1. An introduction to Chinalex

The China Association for Lexicography (Chinalex) was established on October 27, 1992 in Beijing. An Academic Board and the following seven Committees for specific lexicographic fields were set up: Chinese lexicography, Bilingual lexicography, Specialized lexicography, Encyclopaedic lexicography, Editing and Publishing, Computer-aided lexicography, and Theoretical and Historical lexicography.

Cao Xianzhuo was elected as the first President of Chinalex. He was concurrently Deputy Director of the National Language Committee and President of the Institute of Applied Linguistics of the Chinese Academy of Social Science. The following lexicographers and dictionary publishers were elected as Vice-Presidents: Cao Feng (President of Shanghai Lexicographical Publishing House), Wang Yaonan (Professor at Hubei University), Huang Jianhua (President of Guangdong University of Foreign Studies, GDUFS), and Lin Erwei (President of The Commercial Press).

Along with the establishment of Chinalex, a constitution was drawn up and all lexicographic activities were organized to conform to its articles. The President or Vice-Presidents serve for a term of five years. The current President is Cao Guangshun (Chinese Academy of Social Science), and the Vice-Presidents are Yu Dianli (The Commercial Press), Wang Xuming (Language & Culture Press), Liu Qing (China National Committee for Terms in Sciences), Yang Bin (Sichuan Dictionary Publishing House), He Yuanlong (Shanghai Lexicographical Publishing House), Gong Li (Encyclopedia of China Publishing House), Zhang Yihua (GDUFS), and Wei Xiangqing (Nanjing University). The General Secretary is Zhou Hongbo (The Commercial Press), and Chair of the Academic Board is Zhang Yihua.

Over the past 20 years, Chinalex and its subordinate committees have created a platform for Chinese lexicographers to exchange ideas and take part in scholarly activities including lexicographic theory, practice and publication, which helped to enter a new period of rapid development. Many lexicographic institutions were set up, such as the Lexicographic Department of the Chinese Academy of Social Science, the Center for Lexicographical Studies of GDUFS, the Chinese Lexicography Research Center of Ludong University, the Institute of Ancient Books of Hubei University, the Lexicographical Research Institute of Shaanxi Normal University, the Center for Bilingual Lexicography and Bilingual & Bicultural Studies of Xiamen University, the Bilingual Research Center of Nanjing University, the Dictionary Research Institute of Sichuan International Studies University, the Dictionary Research Institute of Heilongjiang University, and the Lexicographical Research Center of The Commercial Press.

Chinalex also sponsors two journals, *Lexicographical Studies* (Cishu Yanjiu) and *Journal of Lexicography in China* (Zhongguo Cishu Xuebao). The former started its publication in 1979, and the latter in 2015.

2. Characteristics of lexicographic practice

All the main dictionary publishing houses in China are members of Chinalex, including The Commercial Press, Foreign Language Teaching and Research Press (FLTRP), Shanghai Lexicographical Publishing House, Shanghai Foreign Language Education Press (SFLEP), Shanghai Translation Publishing House, Sichuan Dictionary Publishing House, and Chongwen Book Company. The most important lexicographic projects, apart from *The Encyclopedia of China*, are all sponsored and published by these publishers, such as *Sources of Chinese Words* (Ci Yuan), *Sea of Chinese Words* (Ci Hai), *Grand Chinese Dictionary* (Hanyu Da Cidian), *Contemporary Chinese Dictionary* (Xiandai Hanyu Cidian), *Grand Dictionary of Chinese Characters* (Hanyu Da Zidian), *Xinhua Chinese Character Dictionary* (Xinhua Zidian), *Xinhua Chinese Dictionary* (Xinhua Cidian), *A New English-Chinese Dictionary* (Yinghua Da Cidian), and *The*

ZHANG Yihua is professor in linguistics and applied linguistics at Guangdong University of Foreign Studies (GDUFS), director of the Center for Lexicographical Studies and member of the Academic Board in GDUFS, vice-president of China Association for Lexicography (Chinalex) and chair of its Academic Board and Bilingual Committee, vice-chair of China National Standardization Committee for Lexicographical Terminology, executive director of the State Committee of Modern Technology for Lexicography, and chief editor of *Journal of Lexicography in China*. He has authored well over a hundred academic publications in lexicography, including papers, works and translations, as well as dictionaries. Among these, *English-Chinese Medical Dictionary* won first prize of the Fifth National Dictionary Award and *Contemporary Lexicography* won the Outstanding Achievement Award of China Colleges and Universities in Scientific Research (Humanities and Social Sciences). His main interests include cognitive linguistics, lexical semantics, lexicography, translation and second language acquisition, and in recent years his research focused on theoretical issues involving the integration of cognitive linguistics and cyber-linguistics theories into lexical and lexicographical researches, computational lexicography, cultural translation, language contact (China English) and foreign-oriented Chinese learning and lexicography.
bilex@mail.gdufs.edu.cn

Chinalex sub-committees, chairs and affiliations

- Academic Board – Zhang Yihua, Guangdong University of Foreign Studies
- Chinese Lexicography – Tan Jinghun, Institute of Linguistics, Chinese Academy of Social Science
- Bilingual Lexicography – Zhang Yihua, Guangdong University of Foreign Studies
- Specialized Lexicography – Peng Weiguo, Shanghai Century Publishing Group
- Encyclopaedias – Gong Li, Encyclopedia of China Publishing House
- Editing & Publishing – Zhou Hongbo, The Commercial Press
- Computer-aided Lexicography – Sun Hongda, Shanghai Lexicographical Publishing House
- Theoretical & Historical Lexicography – Yang Bin, Sichuan Dictionary Publishing House

English-Chinese Dictionary (Yinghan Da Cidian).

Whereas in the English language the basic unit is the word, in Chinese it is the character (字, *zi*). Ancient Chinese consisted only of characters, not words. Along with language evolution, Chinese characters have become very flexible in combination and may be used as fundamental linguistic signs to form words, while many characters maintain the traditional function of encoding semantics in different word classes without any change in form. Thus, we can have both a Chinese character dictionary and a word dictionary, with the following distinctive modern characteristics:

1. the dictionaries cease to function as a tool only to explain hard Chinese characters or words in classic writings, and serve to describe the language in a systematic and comprehensive way;
2. words take the place of Chinese characters and become the main part of the headword list;
3. synchronic description and diachronic explanation are combined (so native-language and foreign-oriented purposes are integrated into one in some bilingual dictionaries);
4. the entry structure is well-established, including word class[1], pronunciation, word sense disambiguation, definitions, examples, collocation, and usage notes.

Recently, reference works of different types are increasingly produced every year.

Table 1 classifies dictionaries published in the last two decades by the three main dictionary publishers in China: Commercial Press, FLTRP and SFLEP. It shows that there are more bilingual dictionaries than monolingual ones, and more dictionaries for foreign language learners than general ones. Nearly all the English monolingual dictionaries originate from British or American publishers, for example, among the 71 English monolingual dictionaries published by SFLEP 41 are from Oxford University Press and 10 are derived from Collins COBUILD. In addition, there is a large number of English-Chinese bilingualised dictionaries, another feature of the local dictionary market that is a sign of

1 Since Chinese words are flexible in use, it has been said that the Chinese language has no word class. Chinese dictionaries for foreign learners began to mark word class in 1995. those for native speakers began to provide it systematically in 2006.

item	category	sub-category	quantity	remark
type	monolingual	English	105	
		Chinese	61	
		other languages	19	
	bilingual (around 87% English)	Foreign Language – Chinese	168	
		Chinese – Foreign Language	112	
		bi-directional	131	
	bilingualized	Chinese-English	87	
		English-Chinese	3	
function	decoding		493	
	encoding		193	
user	native users		197	25 also for foreign users
	foreign/second language learners		512	around 87% English
language coverage	general		366	
	specialized		320	
time coverage media represent-ation	diachronic		46	3 also synchronic
	synchronic		640	
	print		674	
	electronic		12	
	verbal dictionaries		665	
	illustrative dictionaries		21	
user level	beginner		129	
	intermediate		178	6 also for intermediate-advanced
	advanced		379	

Table 1. Classification of contemporary dictionaries by main dictionary publishers in China

the popularity of EFL learning and teaching in China.

It is thus evident that dictionaries in China mainly focus on a synchronic description of language for general-purpose decoding tasks. Fewer encoding dictionaries are found on the market, and most learners' dictionaries are either bilingual or English-Chinese bilingualised ones. Electronic (including online) versions of the main Chinese dictionaries are not available except for a few mobile apps, a serious structural defect in dictionary distribution. However, it was recently announced that the newly revised *Sources of Chinese Words* (3rd edition) will become available in both print and electronic versions (on flash disk and online), and *The Encyclopedia of China* (3rd edition) will also be put online.

In recent decades, along with the increasing zeal for learning Chinese as a foreign language around the world, many learners' dictionaries have been compiled and marketed. The most representative one is *800 words of Contemporary Chinese* (Xiandai Hanyu Babaici, 1980), compiled by the distinguished linguist Li Shuxiang and designed to describe function words and other common words, focusing on the meaning, grammatical pattern, and usages of each lexical unit. Other dictionaries were published successively, such as *Modern Chinese Learner's Dictionary* (Xiandai Hanyu Xuexi Cidian, Sun Quanzhou, 1995), *Usage Dictionary of Modern Chinese Common Words* (Xiandai Hanyu Changyongci Yongfa Cidian, Li Yimin, 1995), *Usage Dictionary of Chinese Common Words* (Hanyu Changyongci Yongfa Cidian, Li Xiaoqi, 1997), *Chinese-English Leaner's Dictionary* (Hanying Shuangjie Cidian, Wang Huan, 1997), *Contemporary Chinese Leaner's Dictionary* (Dangdai Hanyu Xuexi Cidian, Xu Yumin, 2005), and *Commercial Press Learner's Dictionary* (Shangwuguan Xuehanyu Cidian, Lu Jianji, 2007).

A number of Chinese learners' dictionaries for native speakers were published as well. A representative one is *Contemporary Chinese Learner's Dictionary* (Xiandai Hanyu Xuexi Cidian, The Commercial Press, 2010).

3. A new generation of learners' dictionaries

According to Xinhua News Agency, students who received various kinds of higher education in colleges and universities in China numbered 35.59 million by the end of 2014. Furthermore, there were more than 200 million school pupils, including 57.36 million middle school students and 45.27 million high school students. They all learn a foreign language, the majority being English. Chinese higher education attaches special importance to bilingual instruction. In 2001 the Ministry of Education stated that basic and specialized courses for undergraduates should be taught in English or another foreign language. But the students' lack of foreign language proficiency is usually an obstacle to bilingual instruction in specialized courses. The students must turn to learners' dictionaries for unknown lexical information, technical terms and expressions. Therefore, English learners' dictionaries attract much attention from lexicographers, and numerous researches have focused on the theory and practice of English pedagogical lexicography. The Center for Lexicographical Studies of GDUFS has proposed an integrated approach to the EFL learner's dictionary and lexicographic practice, which involves an original design made especially for Chinese learners, including the application and integration of cognitive linguistics and second-language acquisition theories.

Theoretical research has resulted in a dictionary project supported by the National Social Science Fund and SFLTP, called *A New Concept English-Chinese Dictionary for Active Use*. This dictionary features an innovative definition method, which results in a construction-based, meaning-driven, multi-dimensional definition (Goldberg 1995, 2006; Zhang 2006, 2010, 2015b). Event structure, participant/semantic

HUANG Jianhua

deliver *v* 1[VN(prep)]*sb delivers goods, packages, messages, mail or letters etc (to sb or somewhere)*, they take them to that person or place 近义/Syn send, convey, dispatch 传送；递送；投递〔货物；包裹；讯息，邮件或信件等〕(给某人或到某地)…〔例证略〕 反义/Opp receive, get **2** [VN]〖正式〗*sb delivers a statement such as a talk, address or speech, or a presentation, paper, lecture etc*, they make a formal or special account on a particular topic in public, or to a group of people 近义/Syn declare, announce, state, recite 发布，发表〔讲话，演讲，宣言〕；宣读〔论文〕；做〔介绍，讲座〕…反义/Opp conceal, suppress, withhold **3** [VN(Prep)] [VN(prep)]*sb or sth such as an action or event delivers a blow, punch, or kick etc to sb else or a certain part of their body*, they aim at and hit them hard 近义/Syn deal, strike, give 给〔某人〕〔一拳；一脚等〕；猛击[踢]；〔行为或事件〕给〔某人〕沉重打击…反义/Opp receive, suffer, take 派生/Der **delivery** *n*; **deliverer** *n*; **deliverable** *adj*.

fresh *adj* 1[常用于名词前] (of food, vegetables or flowers) recently made, produced, or picked〔食物；蔬菜或花〕新鲜的；新产的；新采的；…**2**[常用于名词前] (of method, style, or looks) completely new, not seen before, and different from of other thing of the same kind 近义 original, novel〔方法；风格；面貌〕新颖的；独特的；…**3** (of sth such as a news, track, act, or memory etc) lately appearing, made or added recently〔消息；线索；行为；记忆等〕新出现的，新近的；新增加的…**4** (of evidence, information and data) newly found or obtained 近义 new〔证据；信息；资料〕新获得的，新发现的；…

Figure 1. A sample entry of *A New Concept English-Chinese Dictionary for Active Use*

Figure 2. Writing interface of a corpus-based dictionary writing system

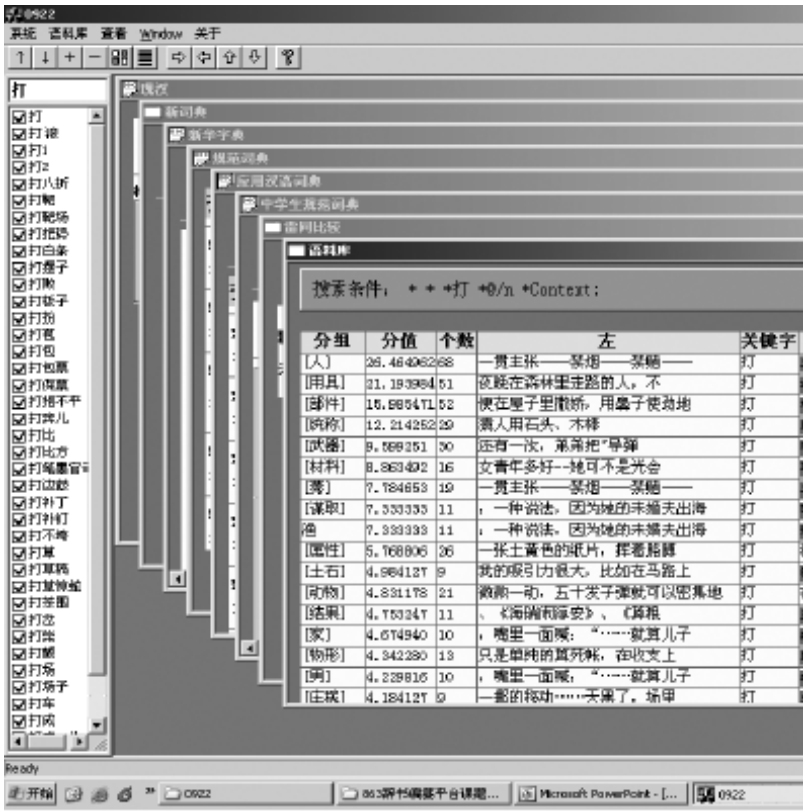


Figure 3. Checking the interface of a corpus-based dictionary writing system

roles, argument structure of related lexico-grammatical constructions, and syntactic-semantic interfaces are clearly shown in the definitional unit as the basis of multi-dimensional meaning representation. Figure 1 presents a sample entry.

A draft version of the dictionary is now completed. Obviously, the style is different from that of existing learners' dictionaries. The definition is bilingualised and firmly based on linguistic studies, and the participant/semantic roles are extracted from a large corpus by means of pattern analysis. Users can thus easily find the necessary morphological, semantic, and syntactic information, as well as co-occurrence patterns and usages of defined words.

4. Computer-aided lexicography

As a cross-disciplinary field of study, computational lexicography has developed into a relatively independent subject through serial researches over a rather long period of time, with a complete set of methodology and research objectives (Zhang 2015a). Recently, Chinese lexicographers have become increasingly aware of the importance of computer technology in lexicography. The main research is focused on dynamic balanced corpus data, semi- or full automation of dictionary writing, formalization of microstructure arrangement, digitization of dictionary media, the *intelligentization* (i.e. having intelligent search and discovery, in China English) of the dictionary query, and integration of multimedia into lexical data representation. The major bodies to have made significant efforts towards these ends – such as corpus building, developing a lexicographical database, or integrating a dictionary writing system – include The Commercial Press, the Center for Lexicographical Studies (GDUFS), SFLTP, and the Institute of Applied Linguistics (under the Ministry of Education).

The Institute of Applied Linguistics developed a corpus-based dictionary writing system that integrates a tagged corpus as well as several mainstream dictionary databases such as *Modern Chinese Dictionary*, *Applied Chinese Language Dictionary*, *Cihai* (Sea of Chinese Words), *Xinhua Chinese Character Dictionary*, and *Verb Usage Dictionary*. Lexicographers can use the system to both write new dictionaries and revise existing ones, as well as to find evidence to support their information. Figure 2 shows the system's interface divided into five parts: the left column displays the headwords; the bottom consists of a dictionary writing template; the top presents the entry preview or display; the middle left column

is used to extract examples from a tagged corpus that can offer information on word segmentation, word class, co-occurrence, grammatical pattern, senses and meaning; the middle right column serves to extract collocations (see also Figure 3).

Some research institutions, for example the Applied Language Institute and the Center for Lexicographical Studies (GDUFS), have proposed a comprehensive approach to electronic lexicography so as to integrate corpus, database, computer-aided compilation and revision, quality control, etc. into one system composed of three parts:

1. **Resources.** Related dictionaries, corpora and language norms and standards, constituting a large general language knowledge base, and serving as supporting information for dictionary writing.
2. **Processing.** Lexical data duplication checking through a conceptual relevance network, i.e., estimating similarities among related dictionaries; lexical conflict checking in a series of dictionaries; lexical normative checking against related linguistic and terminological norms or standards; description and representation of syntactic-semantic interfaces through corpus pattern analysis; and establishing lexical-semantic relations with phonological, morphological and conceptual relevance.
3. **Objects.** The products of the system include the dictionary generation system featuring automatic dictionary production based on a lexicographic database; a checking system as outlined above; an operational interface composed of system management, data-statistics, and multi-property retrieval, the latter including formal, phonetic, and semantic properties; inter-character relevance, sequential value properties; and related resources.

With such a system, almost all operations can be done on a single online-based and computer-aided programme, and all the data necessary for dictionary writing, checking, editing and revising can be made available by a click on the corresponding buttons or control icons.

Cited Dictionaries

Han Zuoli. 2013. *Xinhua Chinese Dictionary (3rd edition) (Xinhua Cidian)*. Beijing: The Commercial Press.

He Jiuying, Wang Ning, et al. 2015. *Sources of Chinese Words (3rd edition) (Ci Yuan)*. Beijing: The Commercial Press.

Lexicographical Department of the Chinese Academy of Social Science. 2012. *Contemporary Chinese Dictionary (6th edition) (Xiandai Hanyu Cidian)*. Beijing: The Commercial Press.

Lexicographical Department of the Chinese Academy of Social Science. 2011. *Xinhua Chinese Character Dictionary (11th edition) (Xinhua Zidian)*. Beijing: The Commercial Press.

Lu Gusun. 2007. *The English-Chinese Dictionary (2nd edition) (Yinghan Da Cidian)*. Shanghai: Shanghai Translation Publishing House.

Luo Zhufeng. 1986-1994. *Grand Chinese Dictionary (Hanyu Da Cidian)*. Shanghai: Shanghai Lexicographical Publishing House.

Xia Zhengnong & Chen Zhili. 2010. *Sea of Chinese Words(6th edition) (Ci Hai)*. Shanghai: Shanghai Lexicographical Publishing House.

Xue Zhongshu. 2010. *Grand Dictionary of Chinese Characters* (*Hanyu Da Zidian*). Chengdu, Wuhan: Sichuan Dictionary Publishing House and Chongwen Book Company.

Zheng Yili, Cao Chengxiu. 2000. *A New English-Chinese Dictionary (3rd edition)* (*Yinghua Da Cidian*). Beijing: The Commercial Press.

References

Goldberg, A. E. 1995. *Construction: A Construction Grammar Approach to Argument Structure*. Chicago: The University Chicago Press.

Goldberg, A. E. 2006. *Constructions at Work: The nature of generalization in language*. Oxford: Oxford University Press.

Zhang, Yihua. 2006. Cognitive semantic structure: A cognitive approach to the essence and structure of bilingual dictionary definitions. *Modern Foreign Languages* (4): 362-369.

Zhang, Yihua. 2010. Cognitive semantics and multidimensional definition for a new generation of bilingual/ bilingualized learner's dictionaries. *Foreign Language Teaching and Research* (5): 374-379.

Zhang, Yihua. 2015a. Computational Lexicography. In Chan, Sin-Wai (ed.). *The Routledge Encyclopedia of Translation Technology*. London & New York: Routledge

Zhang, Yihua. 2015b. *Second Language Acquisition and Learner's Dictionaries*. Beijing: The Commercial Press.

The Center for Lexicographical Studies of GDUFS

The study of lexicography at Guangdong University of Foreign Studies (GDUFS) began in the 1970s by the GDUFS President, Professor Huang Jianhua, a pioneer in modern theoretical lexicography in China, and was followed by Professor Zhang Yihua a leading scholar. In the early 1980's, lexicography became the key study area in the former Institute of International Languages and Cultures and in the mid-1990's the Center for Lexicographical Studies (CLS) was established. From then on, Huang Jianhua, Chen Chuxiang and Zhang Yihua obtained significant achievements in their academic researches, exercising great impact on contemporary lexicography in China. With the development of lexicographical studies and growth of the academic team, CLS became an independent research institution of GDUFS in 2001, and it functions as the headquarters of Chinalex Bilingual Committee, of which Huang and Zhang are successively the former and current chairman.

Lexicographic study at CLS constitutes one of the three well-established research areas of the National Key Research Center for Linguistics and Applied Linguistics and is recognized as National Key Discipline. CLS comprises the following sections: lexicographic research; dictionary compilation; laboratory for computer-aided dictionary compilation; lexicography teaching; and a reference room. It includes 8 full-time faculty members and 10 guest or part-time ones, among them 7 professors and 5 associate professors, including 7 graduate supervisors and three PhD supervisors.

http://cdx.gdufs.edu.cn/

Treatment of entries with Chinese characteristics in English learner's dictionaries: A case study of *Oxford Advanced Learner's Dictionary* 8e

Lixin Xia and Langwei Zhai

XIA Lixin is professor and M.A. candidate supervisor at the Centre for Lexicographical Studies (CLS), Guangdong University of Foreign Studies. He is general secretary of the Chinalex Bilingual Committee, author of more than 20 papers on lexicography and China English, and principal investigator of projects funded by the Ministry of Education or the Guangdong Planning Office of Philosophy and Social Science. His main research interests include lexicography and English varieties. Currently he is visiting scholar at Coventry University.
cdhuiyi@aliyun.com

1. Introduction

Along with the dramatic increase in international exchange between Chinese people and Westerners, more and more words of Chinese origin infiltrate the English language. According to the Global Language Monitor (Radtke 2007), among the 2,000 new words and phrases added to English in 2005, 20% stemmed from Chinese. English learner's dictionary compilers have noticed the phenomenon and adjusted their practice accordingly. However, for various reasons, their treatment of words and expressions with Chinese characteristics requires improvement. A typical case is the *Oxford Advanced Learner's Dictionary, Eighth Edition* (OALD8, 2010), and in this study we examine such entries with Chinese characteristics.

There are two reasons for choosing OALD8 as our study object. The OALD is one of the best-selling English learner's dictionaries worldwide, and the annual sales volume of the bilingual Chinese version of OALD8 reached over one million. Moreover, The bilingualized version (English and Chinese) of OALD8 published by The Commercial Press in Mainland China occupied the first place by sales volume under the category of "English-Chinese/ Chinese-English Dictionaries", according to the statistics of two major online stores (jd.com and amazon.com; data accessed at 21:20 on March 24th, 2016).

2. Research methods of the current study

The 'advanced search' function of the CD-ROM version of OALD8 was used to retrieve all the entries with the tags "originated from Chinese" or "used in the region of China". There were 47 entries with Chinese characteristics which we classified in 8 categories as shown in Table 1.

After the entries were selected and classified they were examined one by one with the aims of identifying possible imperfections in their treatment and making suggestions for improvements if applicable .

3. Analysis of entries with Chinese characteristics in OALD8

The entries with Chinese characteristics were analyzed from the perspectives of headword selection and inclusion, definitions, and labels.

3.1 Headword selection and inclusion

The English words listed in Table 1 have distinctive Chinese characteristics, most of which concern Chinese customs, including terms of sports and entertainment (kung fu, t'ai chi ch'uan, Chinese chequers, mahjong and Cantopop), clothing (cheongsam, samfu) and ways of doing things (feng shui, kowtow, Chinese lantern). The second largest category consist of words and expressions denoting philosophy. Chinese philosophy has a long history, some of

Table 1. List of entries with Chinese characteristics in OALD

category	headword
politics	Maoism, Maoist
economy	renminbi, taipan, yuan
language	Cantonese, Yue, Wu, Xiang, Chinglish, putonghua
philosophy and religion	Dalai Lama, lama, Lamaism, lamasery, Confucian, Taoism, yang, yin
customs	Cantopop, cheongsam, Chinese chequers, Chinese lantern, feng shui, kowtow, kung fu, mahjong, samfu, t'ai chi ch'uan
cuisine	chop suey, chow mein, dim sum, foo yong, hoisin, Peking duck, wok, wonton
craft	china, China-blue, china-clay, kaolin
animal and plant	Chinese cabbage, chow, ginkgo, lapsang souchong, lychee, pak choi

which dates back to over 2,500 years ago, and has profound influence on both Chinese culture and Western philosophy. It is not surprising that words referring to Chinese food constitute the next largest category, as its global popularity and influence have made the names of Chinese dishes enter the English language as loanwords. Table 1 shows that words and expressions with Chinese characteristics in OALD8 cover a wide range of fields from daily life to philosophy, from business to politics, from craft to custom, and from food to language. Moreover, a large proportion of these words and expressions comes from Cantonese, such as *taipan*, *kowtow*, *samfu*, etc. This might be due to the fact that Guangzhou was among the first cities in China that opened its doors to the West in ancient time, and that more immigrants from the Guangdong and Hong Kong areas went to live and work in English-speaking countries and brought their culture and language there.

According to the OALD8 blurb, the total number of headwords in the dictionary is 184,500, which means the number of Chinese-derived words and expressions (47) accounts for barely 0.03% of its entries. Nearly twenty years ago Benson (1997:133) has noted that English learner's dictionaries (ELDs) contain fewer references to China than their larger counterparts such as OED. The OED online version reveals through advanced search 250 entries of Chinese origin, that is 0.04% of the total 600,000 entries. However, as ELDs claim to be specifically designed for foreign learners, they may be expected to include more entries from other varieties than their larger counterparts for native speakers. Moreover, the number of headwords from Chinese is out of proportion to that of headwords derived from other languages, say Japanese. According to Zeng (2005, 2016), the number of headwords originating from Japanese in the *Shorter Oxford English Dictionary* is several times higher than those from Chinese. Last, but not the least, the headwords with Chinese characteristics in OALD8 are mostly from old times, and only a few refer to the current time. Since, as mentioned above, more and more new words from Chinese enter the English language in recent years, ELDs should reflect this language change accordingly.

3.2 The definitions

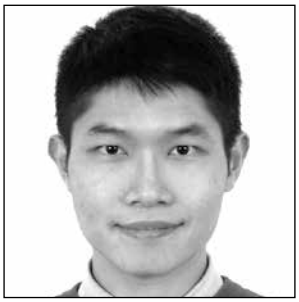
The words and expressions with Chinese characteristics seem to be alien to the OALD8 compilers, as some of them are not defined accurately. Besides, some definitions are too simple or vague and could cause a difficulty in understanding for dictionary users, even for Chinese native speakers.

3.2.1 Regional differences not indicated in the definitions

As a dialect of Putonghua, Cantonese shares the same Chinese character with Putonghua, but has different pronunciations. In many cases, the same word in Cantonese and Putonghua may refer to different referents in the real world, which is liable to lead to confusion.

For instance, the headword *chop suey* is from Cantonese, referring to a kind of mixed food made of meat and vegetables, is defined in OALD8 as "a Chinese-style dish of small pieces of meat fried with vegetables and served with rice". This definition, however, is problematic for native Mandarin Chinese users as this dish in other parts of China is totally different from that in the Cantonese-speaking areas. The pinyin form of *chop suey* in Mandarin Chinese is *zasui*, which refers to a dish of cooked entrails of cattle or sheep (Liu 2009). For speakers of Cantonese and Putonghua, *chop suey* and *zasui* are two different kinds of dishes with different meanings. By reading this definition, speakers of Mandarin Chinese would normally understand *chop suey* to be another kind of dish rather than *zasui*.

Another example is *cheongsam*, which is defined as "a straight, tightly fitting silk dress with a high neck and short sleeves and an opening at the bottom on each side, worn by women from China and Indonesia". While this word form is in Cantonese, the dress itself originates from Shanghai and was made fashionable by upper-class women at the beginning of the 20th century, referring to an exclusively traditional gown also known in Mandarin Chinese as *qipao*. For Cantonese speakers in the areas of Guangzhou, *cheongsam* usually constitutes a jacket with long sleeves, not necessarily a long dress covering the whole body. In Hong Kong, *cheongsam* is a dress for both women and men. Besides, the pinyin form of *cheongsam* is *changshan* in Mandarin Chinese, which denotes exclusively a piece of clothing for men: a long and loose-fitting piece of clothing that covers all of one's body and reaches the ground, worn especially by educated men in ancient China. It was a sign of rank, or at least of literacy, because at that time poor people were mostly illiterate and could not afford a piece of *cheongsam* (Xia 2015). It seems that OALD8 adopted the signifier of the concept in Cantonese and the signified object from Shanghai areas. No regional uses were shown in the definition. As a result, speakers of Mandarin Chinese will have difficulty in understanding the definition. Moreover, an English speaker travelling in China will also feel puzzled when he orders a *chop suey* or *cheongsam*



ZHAI Langwei is M.A. candidate at the Centre for Lexicographical Studies, Guangdong University of Foreign Studies. His main research interests include lexicography and cognitive semantics.
longwaychai@hotmail.com

Dictionnaire de Poche
Français-Chinois
Chinois-Français
Commercial Press
International
Beijing, China
March 2016
592 pages, 75x110x25 mm
PVC
ISBN: 978-7-5176-0163-0
RMB 29.80
http://www.cp.com.cn/

Published in cooperation
with K DICTIONARIES

but is served with or given a different dish or dress. One may argue that OALD8 describes these Chinese words in the English language, not their use in Chinese. The description of the English usage of these words, however, is not correct because their referents are not the ones they refer to in their original use in Chinese. As a dictionary is by nature both descriptive and prescriptive, it should inform its users of the right use of a foreign word in order to avoid possible misconception.

3.2.2 Narrower or wider meanings or extensions

Due to the cultural difference, the same concept may have different meanings or extensions. For example, *Maosim* is defined as "the ideas of the 20th century Chinese communist leader Mao Zedong", a political term that denotes the concept of Mao Zedong's thought as it is termed in China. It was formerly believed to be introduced and developed solely by Mao Zedong, but following the open policy adopted in China since 1978, the term was officially redefined in the Communist Party Committee's Constitution as "Marxism-Leninism applied in a Chinese context", synthesized by Mao Zedong and China's "first-generation leaders" (Qi 2010). According to the official definition, it is the fruit of the collective wisdom of Mao Zedong together with other communist leaders of the first-generation from the 1920's until Mao's death in 1976. The current definition in OALD8 thus has a narrower sense as it is limited to Mao's personal political theories.

The headword *dim sum* is defined in OALD8 as "a Chinese dish or meal consisting of small pieces of food wrapped in sheets of dough". As a matter of fact, besides this sense *dim sum* refers also to Chinese sponge cakes, vegetables wrapped in dried bean milk cream in tight rolls, beef or pork meatballs, and so on. The pinyin form of *dim sum* is *dianxin*, which, in Chinese culture, refers to snacks, light refreshments or desserts that are served, often with tea, in small portions. The definition is thus incomplete in that it only covers one kind of *dim sum*.

On the other hand, the definition of *ginkgo* in OALD8 has a wider meaning that can encompass many other trees as well: "a Chinese tree with yellow flowers". In biological terms, *ginkgo* refers to the plants of the ginkgo genus, the only living member of the gymnosperm family Ginkgoaceae. It has great biological and economic value in that it has a number of primitive features and its fruits can be used in food and medicine. The main characteristics of *ginkgo*are its fan-shaped leaves and yellow flowers.

3.2.3 Denotative meanings not defined

As common practice, the denotative or literal meanings of a headword should be first included and explained, then the extended meanings can be further illustrated. Otherwise the additional meanings would seem to come from nowhere.

For example, *kowtow* is defined in OALD8 as "to show sb in authority too much respect and be too willing to obey them". Its denotative meaning in Chinese is to kneel and touch the ground with the forehead. It originates from the rite of *dunshou*, which consists of three steps – namely, keeling, bending over the body, and touching the ground with the forehead – and is the solemn rite of an inferior to a superior as formerly done in China. The metaphorical meaning of *kowtow* is to show someone, especially one's superior, deep respect, worship, or submission. OALD8 has adopted only the metaphorical meaning of *kowtow* and ignores its literal meaning.

3.2.4 Definiens not included in the definienda

ELDs claim to use a limited defining vocabulary to define all the headwords, and all the definiens are included in the dictionary as definienda themselves. However, the definition of *Taosim* has *Lao-tzu*, which is not included in OALD8: "a Chinese philosophy based on the writings of Lao-tzu". Lao Tzu, an ancient Chinese philosopher, was traditionally regarded as the author of the holy book *Tao Te Ching* and the founder of Taoism. In the book of *Lao Tzu*, Tao is considered as the basic source and supreme law of everything in the universe. The followers of Taoist teaching should stick to the state of vacancy and stillness mentally and physically to understand the nature of Tao. The dictionary includes the entries *Taoism* and *Taoist*, but not *Lao Tzu* or *Tao*, which may cause problems to users who are not familiar with the concept of Taoism. It is common practice for British general language dictionaries not to include proper names, which might result in such shortcomings, whereas American English dictionaries tend to include proper names.

3.3 The glosses and labels

Labels are used in dictionaries to remind users of additional meaning and usage for a lemma. As headwords of entries with Chinese characteristics have specific cultural connotations, it is advisable to illustrate them by way of labels or notes.

The headword *taipan* in OALD8 is defined as "a foreign person who is in charge of a business in China". However, this is an informal term used during the 19th and early 20th century, and is passing out of current

use (Xia 2015). Therefore, a register label, such as *old-fashioned*, should be added to guard users against misusing it. The label means that the word is no longer used, but its counterparts in the real world exist.

The headword *Lamaism* is defined as "Tibetan Buddhism" by way of a synonymous paraphrase. The Chinese equivalent is *lama jiao*, an informal term for Tibetan Buddhism, where *jiao* means a religion. As a matter of fact, according to Tibetan Buddhists and researchers, *lama jiao* is offensive, which might mislead language users to regard it as an independent religion only worshiping the lamas instead of Buddha, or even creating their doctrines from nowhere but the teaching of Buddha (Lopezth Jr., 1999: 6). It, therefore, is recommended that a note be added to warn the dictionary users against misusing it.

OALD8 does provide a gloss for the entry *t'ai chi ch'uan*, defined as **"**(also t'ai chi) a Chinese system of exercises consisting of sets of very slow controlled movements". The entry thus lists *t'ai chi* as a variant. The terms *t'ai chi ch'uan* and *t'ai chi* are closely related in form and content but denote two different concepts, in which the former refers to a Chinese martial art characteristic of slow movements, and the latter to the ancient Chinese philosophy. According to the first half of *Xi Ci* in the *Book of Change*, the source of change is *t'ai chi*, which produces *Yin* and *Yang*. (Editing 2000: 340) In other words, *t'ai chi* is the source of everything and also the essential factor and condition for the change of everything. *T'ai chi ch'uan* is based on the philosophy of *t'ai chi*. Although *t'ai chi* is often used as the shortened form of *t'ai chi ch'uan* both in the West and in China, not denoting the difference might cause confusion to the dictionary users.

Conclusion

From our analysis it can be concluded that lemmas with Chinese characteristics in OALD8 have not been sufficiently well treated, although the dictionary, on the whole, is of high quality. A disproportionate number of Chinese-derived entries has flaws in the definitions and representations. Specifically, among 47 lemmas with Chinese characteristics, 10 have some flaws, which makes up 21%. The first cause for such flaw seems to be that they are not accorded the same status as other English lemmas, such as Japanese-derived ones. Another cause might be that these words have rich cultural connotations that make their compilation difficult for ELD lexicographers. It is understandable that they are not familiar with words and expressions with Chinese characteristics, but neither are the dictionary users. Therefore, we have all the more reasons to clarify such occurrences in the dictionary and define them correctly and accurately.

Acknowledgements

This work was supported by the Humanities and Social Science Research Funding of the Ministry of Education of the People's Republic of China under Grant No. 15YJA740048 (China English or Chinglish? A Study Based on China English Corpus), China Scholarship Council under Grant No. 201508440380, and the Post-graduate Division of Guangdong University of Foreign Studies under Grant No. 14GWCXXM-45.

References

Benson P. 1997. English Dictionaries in Asia: Asia in English Dictionaries. In M.LS. Bautista (ed.), *English is an Asian Language: The Philippine Context*. Sydney: The Macquarie Library, 125-140.

Editing Committee of the Thirteen Confucian Classics with Notes and Commentaries. 2000. *The Book of Change*. Beijing: Peking University Press.

Liu, H. 2009. Chop suey as imagined authentic Chinese food: The culinary identity of Chinese restaurants in the United States. *Journal of Transnational American Studies*, 1(1).

Lopez Jr., D.S. 1999. *Prisoners of Shangri-La: Tibetan Buddhism and the West*. Chicago: University Of Chicago Press.

OED online. *Oxford English Dictionary* online. http://www.oed.com/. Accessed 24 March 2016.

Qi, J. 2010. *Five Factors Influencing China's Foreign Policies*. Beijing: The Central Compilation and Translation Bureau.

Radtke. O.L. 2007. *Chinglish*. Layton: Gibbs Smith Publisher.

Xia, L. 2015. The Corpora of China English: Implications for an English-Chinese Learner's Dictionary. Paper presented at *Australex 2015* held at Auckland University on 18-22 Nov. 2015.

Zeng, T. 2005. Direction of the English translation of Chinese-culture-loaded words: A case study of the Shorter Oxford English Dictionary, 5th edition. *Journal of Guangdong University of Foreign Studies*, Vol. 16 (Supplement): 74-77.

Zeng, T. 2016. Why not translate the Chinese-culture-loaded words by transliteration?. *Nanfang Weekend*, 2016.2.24.

Random House Webster's College Dictionary
New Edition
Commercial Press International
Beijing, China
March 2016
1960 pages, 185x265x70 mm
Hardbound
ISBN: 978-7-5176-0191-3
RMB 298.00
http://www.cp.com.cn/

Published in cooperation with K DICTIONARIES

Lexicography at the Society for Danish Language and Literature

Lars Trap-Jensen

Lars Trap-Jensen has an educational background in general linguistics, Greenlandic, and social studies from Aarhus University, with an MPhil in linguistics from Cambridge University. He was a lecturer in Danish language at the universities of Basel and Zürich. Since 1994 he has been working as a lexicographer at the Society for Danish Language and Literature, Copenhagen, and since 2003 as the managing editor of *The Danish Dictionary* and the Society's dictionary site ordnet.dk. Other projects include the digitization of *Dictionary of the Danish Language* and the development of the Danish wordnet (DanNet) and *The Danish Thesaurus*. Currently he is President of Euralex and involved with the establishment of Globalex.
ltj@dsl.dk

Introduction

The making of dictionaries has been an ongoing activity at the Society for Danish Language and Literature (Det Danske Sprog- og Litteraqturselskab, DSL) for just over one hundred years. In 1915, the Society was encouraged to take the responsibility for the compilation of an ambitious dictionary project, *Ordbog over det danske Sprog* (*Dictionary of the Danish Language*). The outline of this dictionary had already been sketched over the previous decades by Verner Dahlerup, a professor of Nordic philology at the University of Copenhagen. His inspiration came from the grand projects initiated for German, English, Dutch and Swedish, but when he signed a contract to compile the dictionary, in 1901, the plan was for a more modest publication, twice the size of the standard dictionary of the time, Christian Molbech's two-volume *Dansk Ordbog* (*Danish Dictionary*), but still in the format of a concise dictionary. In the following years, he had to revise his plans, now aiming at an estimated 8-12 volumes. Eventually, Dahlerup realized that the task was beyond the working capacity of a single man and thus, in May 1915, he turned to the Society, which had been established only four years earlier.

Foundation and objectives of the Society

The Society was lead by a remarkable woman, Lis Jacobsen, who had also been the driving force in founding the institution. Jacobsen (nee Rubin) hailed from a Jewish family and was the daughter of the national bank governor. She had been the first woman to obtain a doctorate in Nordic philology, and only the seventh female doctor in the country at the time, with a dissertation on the earliest manifestations of the Danish language. About one year later, on 29 March 1911, she gave a lecture on the "means and ends of Danish linguistic research", arranged by the Society for German Philology. She had imagined her dissertation to be just the opening volume of a more ambitious documentary work of the entire history of the Danish language, but had found herself forced to discontinue her work due to the lack of satisfactory source material. Scholarly editions of the source material were scarce and their systematic studies correspondingly few. In her lecture, she stressed the need for both, concluding: "The means to obtain this are twofold: we need *money* and we need *labour*. Money to publish the source material and labour to process it". Present in the audience that day was Kristian Erslev, a professor of history and one of the pioneers of historical criticism and the modern science of history. More importantly in this connection, however, he was also head of the university at the time and in addition a prominent member of the Carlsberg Foundation, later to become its President. He envisioned the perspectives of Jacobsen's message and realized that an institutional framework was needed. His advice to her was to form an editorial society: "If you can provide the labour, I will provide the money". Only one month later, on 29 April 1911, the Society for Danish Language and Literature became a reality.

With that, several important traditions had been established: the goal of the Society was to create scholarly editions of the source material for the study of Danish language and literature through all historical periods and, equally important, a long-term cooperation had been set up with the Carlsberg Foundation as an important and generous sponsor of the Society's activities.

Private foundations as culture bearers

Today, the Society for Danish Language and Literature functions as an independent scholarly institution receiving annual funding from the Ministry of Culture. This covers the administration and operation of services, whereas most scholarly activities are sponsored by external donors for specific projects. Among these is the Carlsberg Foundation, owner of the Carlsberg Group and the world's third largest brewing company. Established in 1876, this industrial and commercial foundation is among the oldest of its kind worldwide. The statutes stipulate that part of the company's profit must be channeled back to society through donations to science and culture, and in this way, Carlsberg has left its mark on many aspects of Danish society. The same is true for a number of other commercial foundations that have financed or co-financed lexicographic projects within the Society: a Swedish-Danish dictionary was sponsored by the foundation owned by A.P. Moller–Maersk Group, the largest company in Denmark and a world

leading container ship operator; the Velux Foundation, producer of windows and skylights, sponsored the digitization of the Old-Danish Dictionary archive, and the Augustinus Foundation, majority share holder in the Scandinavian Tobacco Company, recently gave a donation to *Den Danske Ordbog* (*The Danish Dictionary*). In addtion to the private foundations, projects may also receive donations from special allocations provided for in the Finance Act. The two large monolingual dictionaries, *Ordbog over det danske Sprog* and *Den Danske Ordbog* were both mainly sponsored jointly by the Carlsberg Foundation and the Ministry of Culture.

The *Dictionary of the Danish Language*

Ordbog over det danske Sprog marks a turning point in Danish lexicography which, prior to its publication, had been dominated by the prescriptivism inherited from the tradition of the French Academy. Dictionaries of the 19th century were preoccupied with educating the public, more specifically by protecting it from what was considered bad linguistic influence. The dictionaries should only contain, according to Molbech in Dahlerup's reading, "good" words, "the most beautiful flowers of the language". For a word, it was a mark of honour to be included in the dictionary, much in the same way as it is an honour for a work of art to feature in the nation's art collection. Dahlerup broke away from this tradition and insisted on greater professionalism, declaring: "I cannot ask first of all: "*should* this or that word be used?", but rather: "*is* it used, or *has it been* used?"; if this is the case, I include the word in so far as considerations of space permit" (Dahlerup, 1907).

Where the editors of the 19th century dictionaries had been generalists with mainly educationalist concerns, the editors of *Ordbog over det danske Sprog* in contrast were specialized philologists with intimate knowledge of the language described. At the centre of their work lay a large collection of notes with excerpts from a range of texts. Even if the technology was different and the texts dominaed by exemplary literary and journalistic efforts, the methodology used was not much different in nature from the modern corpus-based approach of descriptive lexicography: from the underlying language material they extracted whatever facts of form, meaning and word patterns they could observe about the linguistic units.

With more than 225,000 entries, *Ordbog over det danske Sprog* is the largest monolingual dictionary compiled for Danish. It is, admittedly, not as comprehensive as its sister dictionaries in Germany, Sweden, the UK and the Netherlands, but the 28 volumes were completed within 40 years, later increased by 5 supplementary volumes, and even to this day, it is, for its size, quite uniform and easy to read and use.

The *Danish Dictionary* and other dictionaries

Its successor, *Den Danske Ordbog*, was launched in 1991 as the first, and so far only, corpus-based monolingual dictionary for Danish. Originally conceived as a paper dictionary (6 volumes, 2002-2005), it has seen its greatest success as an online dictionary, with nearly 100,000 visitors on a normal day (May 2016). It has been online since 2009 on the Society's modern dictionary website (http//ordnet.dk/) along with a digital version of *Ordbog over det danske Sprog*. Unlike the historical *Ordbog over det danske Sprog*, *Den Danske Ordbog* is being updated on a regular basis.

In line with the statutes, the Society aims to provide dictionary coverage of the Danish language across all historical periods. A dictionary of Old Danish, covering the period 1100-1515, has been underway for more than 60 years and is now drawing near its conclusion. The period between Old Danish and Modern Danish is the weakest in terms of coverage, but a series of mainly bilingual Latin-Danish and Danish-Latin glossaries from the Danish Renaissance have been published, and just a few years ago the Society was able to publish for the first time ever the earliest comprehensive dictionary of Danish, compiled around 1700 and describing the language in the latter half of the 17th century. Until then, Matthias Moth's dictionary had only existed as a manuscript in the Royal Library in Copenhagen, but a long-cherished wish for publication was at last made possible through a gift donation from the Carlsberg Foundation in connection with the Society's 100th anniversary in 2011.

In addition to these comprehensive dictionaries, the modern period is also represented by the recent publication of a Danish thesaurus, *Den Danske Begrebsordbog*, as well as two bilingual dictionaries with Swedish and Icelandic as the respective source languages. Furthermore, the Society has recently retro-digitized and published online some of the more important Danish dictionaries, either compiled by the Society itself or by others, taking advantage of the experience gained from the retro-digitization of *Ordbog over det danske Sprog,* by means of double-keying following the model of

Verner Dahlerup

Lis Jacobsen

The DSL building in Copenhagen

Deutsches Wörterbuch of the Brothers Grimm in Germany. Digitized dictionaries of this kind include the *Holberg-Ordbog* (*Holberg Dictionary*), a dictionary of the complete works by the Danish-Norwegian author Ludvig Holberg (1684-1754) published in 5 volumes in 1981-1988, and *Meyer's Fremmedordbog* (*Meyer's Dictionary of Loan Words*), based on the 8th edition from 1924. Most recently, the *Ordbog til det ældre danske Sprog* (*Dictionary of Older Danish*), edited and published by Otto Kalkar in 1881-1918, is being digitized as part of an ongoing project examining the Danish language and literature in the Middle Ages.

References

Andersson, H. 2006. ODS – træk af en historisk ordbogs historie (The Dictionary of the Danish Language – Outlines of the history of a historical dictionary). In Bergenholtz H. and Malmgren S.-G. (eds.), *LexicoNordica* 13, 25-39. Gothenburg.

Dahlerup, V. 1907. Principer for ordbogsarbejde (Principles for Dictionary Work). In Kristensen M. and Olrik A. (eds.), *Danske Studier*, pp. 65–78, Copenhagen.

Carlsbergfondet og "Sprogmindesmærkerne". 2015. News post about the Carlsberg Foundation and the Language "Monuments", accessed May 2016 (http://www.carlsbergfondet.dk/da/Skjulte-sider/Skjulte-artikler/Danske-versioner-af-forskningsprojekter/Sprogmindesmaerkerne/).

Trap-Jensen, L. 2010. Den Danske Ordbog på nettet. http://sprogmuseet.dk/ord/den-danske-ordbog-pa-nettet/.

Trap-Jensen, L. 2012. Ordbog over det danske Sprog. http://sprogmuseet.dk/ord/ordbog-over-det-danske-sprog/.

Towards Peoplex

Ilan Kernerman

I was thrilled to take part in the Dictionaries in Asia conference and the inauguration of Asialex. The need for a forum of this kind has long been felt, and the event lived up to expectations.

It might seem strange no such framework existed so far, since Asia was the cradle for dictionary-making thousands of years ago, and its lexicographic tradition has flourished through the ages to modern times. The 20th century's prominent milestones in pedagogical lexicography stem from the work of Michael West in India and A.S. Hornby in Japan. Some of the world's finest dictionaries are made in Japan and its neighbors, as well as valuable research carried out, but these are little known of elsewhere.

In addition to economic-political factors, this lack may be mainly due to Asia's inherent diversity, not being a homogenous entity of any sort. Linguistically, unlike most European tongues that pertain to the Indo-European family,.Asian languages share no common background, apart from being human.

That natural human link is true just as well for the entire world. Asia can project a microcosm of it and, thus, establishing Asialex is a significant step toward forming a global lexicographical constellation.

A future GLOBALEX (or Unilex, in the words of Tom McArthur) concerns globalization and co-existence in multilingual societies, English as the international lingua franca, localized Englishes, effects on the mother tongues, etc, as well as repercussions from hi-tech and tele-communication, online interactivity and automatic translations, *Dictionizers* and *Quicktionaries*, and so on.

This forthcoming forum should not replace national or regional LEX's, but accommodate the varied issues. As such, geography is no sound base for its foundation, nor for the soon-to-come dictionaries that will hardly be what we imagine now.

Beyond countries and behind computers there are people. First of all, and after all. People are the most common denominator for lexicography all over the world.

(Reprinted with slight amendments from *KDN* 5, 1997.)

Linked data in lexicography

Julia Bosque-Gil, Jorge Gracia and Asunción Gómez-Pérez

1. Introduction

The notions of *linked data* (LD) and *Web of Data* are increasingly gaining ground in digital humanities, linguistics, biomedicine, e-science, data journalism, etc. and lexicography is not staying behind. The LD paradigm meets the need to link isolated pieces of information which were in their own proprietary formats and were previously hard to discover and integrate. The term actually refers to a "set of best practices for exposing, sharing, and connecting data on the Web" (Bizer et al. 2009). In order to create LD there is a set of requirements to fulfill, among them, the use of Unique Resource Identifiers (URIs) and the establishment of links to other resources. The Resource Description Framework (RDF)[1] is the formal backbone giving support to this network of interlinked resources and allowing for the definition of *triplets* or statements of the form *subject-predicate-object*, where *subject* and *object* are resources and the *predicate* is the edge or *property* connecting the nodes. The result is a vast graph whose nodes can be practically anything, including lexical units, and this is where lexicography comes into play.

The work in models for the representation of linguistic information as LD (McCrae et al. 2012), as well as in best practices and guidelines for the conversion of mono- and multilingual language resources[2] has been continuous in recent years. The benefits that LD brings to lexicography have been already pointed out in recent works related to the conversion of bilingual and multilingual dictionaries as LD (e.g. Gracia 2015, Klimek and Brümmer 2015, Bosque-Gil et al. 2016) and etymological and dialectal dictionaries (Declerck et al. 2015, among others), as well as in recent initiatives and international projects that have embraced the use of semantic technologies[3] and in current e-lexicography work (McCracken 2015). The main advantages are the semantic and syntactic interoperability provided by RDF and linguistic vocabularies (LexInfo[4] or GOLD[5]), which enables the integration, exchange, and enrichment of lexicographic data among different resources, the reusability of the whole resource, which in turn prevents lexicographers from "re-inventing the wheel" in potential future projects, improved data visualization and querying, resource sustainability (Wandl-Vogt 2015), and easy discovery thanks to metadata repositories.[6]

In this context, this paper seeks to present, on the basis of our experience in the conversion of lexicographic data to LD[7], our reflections on the implications of converting lexical data to LD, drawing special attention to the advantages it offers from the eyes of a lexicographer or a linguist outside the realm of the Semantic Web, but as part of a discipline which can be already considered part of information science (Fuertes-Olivera and Bergenholtz 2011). Our goal is therefore twofold: to place LD in the context of lexicographic work in lexical networks, and to bring its benefits closer to the lexicographer so she can consider it a basis for future endeavours. To this end, we will first provide a brief overview of the work on the representation of lexical information as graphs outside the context of the Semantic Web with focus on WordNet (Miller 1995, Fellbaum 1998) and Polguère's lexical systems (Polguère 2012, 2014) implemented in the French Lexical Network (Gader et al. 2012). Then, we will dwell on the practical advantages of LD for representing both the macro- and the microstructure of a lexicon.

2. Lexical data as a graph

Modeling lexical information as a graph is not a novel notion coming from LD.

1 https://www.w3.org/TR/rdf11-primer/
2 https://www.w3.org/community/bpmlod/
3 Such as the ENel cost action (http://www.cost.eu/COST_Actions/isch/IS1305/) or the LIDER (http://lider-project.eu/) and LDL4HELTA (http://www.eurekanetwork.org/project/id/9898/) projects
4 http://www.lexinfo.net/ontology/2.0/lexinfo/
5 http://www.linguistics-ontology.org/
6 linghub.org, http://metashare.elda.org/
7 From October 2015 to February 2016, the Ontology Engineering Group at UPM worked on the development of a linguistic linked data prototype for K Dictionaries and Semantic Web Company as part of their LDL4HELTA project, and, more specifically, on the transformation to RDF of the Spanish dataset of K Dictionaries.

Julia Bosque-Gil is a PhD student at the Ontology Engineering Group, Universidad Politécnica de Madrid. She holds a B.A. in German Linguistics and English from the Humboldt University of Berlin and an M.A. in Computational Linguistics from Brandeis University, Waltham. Her interests include the lexicon-ontology and the syntax-semantics interfaces, the relations between the lexicon and syntax, semantic annotation and the representation of (multilingual) language resources as linked data. She has been working in the conversion of multilingual terminologies to RDF as part of the LIDER project and in the modeling of lexicographic data as linked data in collaboration with K Dictionaries and Semantic Web Company. For her PhD thesis she is investigating the use of linguistic linked data for research in linguistics.
jbosque@fi.upm.es

Jorge Gracia is a post-doctoral researcher at the Ontology Engineering Group, Universidad Politécnica de Madrid, Spain. He got his PhD in Computer Science at University of Zaragoza in 2009, with a thesis about heterogeneity issues on the Semantic Web. His current research interests include multilingualism and linked data, linguistic linked data, and cross-lingual matching and information access on the Semantic Web. Currently he is exploring how to move language resources (lexica, dictionaries, corpora, etc) from their data silos into the multilingual Web of Data and make them interoperable, in order to support a future generation of linked data-aware NLP tools. http://jogracia.url.ph/web/

WordNet already set a precedent (Miller 1995, Fellbaum 1998) as a graph-based lexico-semantic database where nodes represent the concepts (synsets or sets of cognitive synonyms) and hyponymy, meronymy and antonymy constitute the relations that link them together. Furthermore, there are other efforts in lexicography that emerge from a conception of the natural language lexicon as a network of entries rather than a list, which is what the organization of conventional dictionaries looks like. The entries are then viewed as part of a language system of related lexical elements. Polguère's notion of a lexical system, implemented in the framework of the French Lexical Network project, falls into this category. However, in contrast to the projects developed in lexical semantics, linguistic linked data (LLD) and the models proposed for converting resources into them (*lemon*[8], SKOS-XL[9], LIR (Montiel-Ponsoda et al. 2008)), do not arise as initiatives to model the (mental) natural language lexicon, nor make such claim, even though they entail the use of classes and properties such as *lexical entry*, *sense*, *lexical concept*, *syntactic frame*, *lexical form* or *definition*. LD emerges as a technological means to better represent, share, integrate and discover linguistic knowledge scattered over the Web and its underlying RDF formalism is not conceived from a theoretical perspective as an alternative to structure mental lexical information. Nonetheless, knowing the direction into which lexical semantics and lexicography move, as well as the similarities between the representations suggested there and those proposed from an LD perspective, will help us in building bridges for collaboration between experts from both sides. LD, as best practices for data representation, should be compatible with the representation of any lexical network, even though this implies the extension of vocabularies currently available on the so-called linguistic linked open data (LLOD) cloud[10] or models to encode all the data that the theory on which the resource is based addresses.

An analysis of what modeling the lexicon as a graph in WordNet entails and which needs are met is given in Polguère (2014) and McCracken (2015): lexical entries were previously analyzed and presented independently one from another and a novel approach reflecting what the structure of the mental lexicon might resemble was called for.[11] WordNet falls under the category of ontology-based lexical network (Polguère 2014: 3), i.e. a network of lexical units with an ontology as backbone, including word senses arranged in a hierarchy and related by synonymy, hyponymy and meronymy relations. It is worth mentioning that LLD relies on linguistic ontologies or vocabularies, but the creation of an ontology of word senses or concepts is actually optional and it is not a required step in order to publish LD. Accordingly, we can state that the entry *enthusiasm* in an English lexicon has as the part-of-speech `lexinfo:noun`, which is defined along with, for instance, `lexinfo:reflexivePersonalPronoun`, as an individual of type `lexinfo:PartOfSpeech` in the linguistic ontology LexInfo.[12] We are thus linking two resources without establishing the concept denoted by *enthusiasm* in any hierarchy (e.g. as a child of *feeling*). LD resources such as BabelNet[13], DBpedia[14], and WordNet RDF (McCrae et al. 2014) have an underlying ontology, but this is not implied in the conversion of every resource to LLD. In relation to this, LLD builds upon the notion of *semantics by reference* (McCrae et al. 2012): the meaning of a word and the word itself (the *signifier -- signified* opposition) are separated in two different layers, with `ontolex:LexicalEntry` and `skos:Concept` respectively, and the relation between the two is "reified" in a class that aims at encoding *a sense* (`ontolex:LexicalSense`). All the linguistic information pertaining to the word itself or to the use of that word with that specific meaning is separated from the actual meaning, which, ideally, is language independent. Hierarchic conceptual relations would be established, if they are, at the level of the concept.

Polguère places lexical systems on the other side of the balance: they are lexical networks that are not ontology-based. Lexical systems are conceived with the relations among the lexical elements as focus and relegate to the background the classification of units or property inheritance (Polguère 2014: 3). A key aspect of lexical systems is that relations are not limited to synonymy, hyponymy, etc. but they include

8 http://www.lemon-model.net/lemon/
9 https://www.w3.org/TR/skos-reference/skos-xl.html/
10 http:/linguistic-lod.org/llod-cloud/
11 However, most lexical semantics research addresses different aspects with different levels of granularity, but it does not analyze all word types and the semantic structure of the lexicon as a whole (Swanepoel 1994)
12 http://www.lexinfo.net/ontology/2.0/lexinfo/
13 http://babelnet.org/
14 http://wiki.dbpedia.org/

paradigmatic and syntagmatic relations drawn from the set of lexical functions of the Meaning Text Theory (Mel'čuk 1996). The result is a multi-dimensional graph with a wide range of relations linking the nodes (lexical elements), which instantly brings RDF to mind. There are two important points to bear in mind when comparing lexical systems with resources migrated to LLD: first, the nodes in lexical systems are already "disambiguated", each node represents one specific meaning of the lexical unit at hand (Polguère 2014: 5). The closest counterpart we have in RDF is `ontolex:LexicalSense`, which is a unique relation between a word and a meaning. Secondly, the nodes in a lexical system are not atomic and each one records the information we would find in a lexicographic article. Grammatical information, semantic label, syntactic government pattern (collocations are implemented by edges), etc. are stored inside the node. In LLD, some of these data would be linked to the entry at hand or to one of its `ontolex:LexicalSense`(s) by means of specific properties (edges) and elements available in linguistic vocabularies, identifiable with their own URIs: lexical entries, word forms, senses, part of speech tags, gender, number, subcategorization, etc. To see which entries are related, the SPARQL query language[15] allows to perform queries on the graph and trace the connections between ontolex lexical senses or ontolex lexical entries. LexInfo, *lemon-ontolex*, SKOS, GOLD, etc. already provide a high number of relations, which can be extended with new ones or new vocabularies can be created as needed.

In sum, the idea of representing lexical information as a graph is not new, and LD are not presented as a novelty in this regard. However, they allow for the implementation of networks or the integration of already available ones on the basis of a homogenous format. Thus LD meet the need for linking lexical elements that were previously isolated by using sets of relations and elements that are defined externally and can be extended as required, relying or not on an underlying ontology of word senses. This does not mean that LD is equivalent to any of the efforts mentioned above or forms a better option to the structures in which they are implemented, and, as said, it does not make claims on the structure of our mental lexicon. RDF is, however, a model to represent data worth taking into consideration for lexicographic projects aiming at the creation of lexical networks because it provides a basis for their implementation while retaining all the benefits related to interoperability, visibility and NLP-services compliance.

All in all, the LD paradigm is agnostic with respect to the different theories in modern lexicography, and it poses a number of tangible benefits that we enumerate in the following sections.

3. Benefits of a lexicon in linked data: macro-structure

Having placed LD in context, what are the actual benefits of creating or converting a lexicon to LD? The most evident advantage is that LD enable the integration with other external resources thanks to the semantic and syntactic interoperability achieved by the use of RDF and linguistic ontologies. Besides this fact and focusing on the lexicon itself, we have identified the following benefits in the course of our work towards the migration of language resources to LLD, some of them also highlighted in the literature.

Firstly, the entries of a dictionary become internally reusable (Klimek and Brümmer 2015) thanks to their URIs. This does not seem novel given that entries might already have numeric identifiers to point to each other, but the choice of transparent URIs, i.e, human-readable, which reflect the semantic content, and a suitable URI naming strategy play a crucial role (Bosque-Gil et al. 2016): the editor of a dictionary entry will be able to refer to another entry without the need to know its identifier in advance. Following this, the entry `:lexiconEN/risk-n` can be linked to `:lexiconEN/risky-adj` through a relation of morphologic derivation without the need of an identifier. If, later on, the noun *risk* occurs as an entry in another dictionary of the same or a different family of dictionaries, the information can be integrated in a straightforward manner without relying on dictionary-dependant numeric IDs.

This in turn relates to a second advantage: we no longer depend on the order of appearance of lexical entries or senses in cross-references, which is usually indicated by a superscript in numeric form in printed or electronic format, e.g: *bow*^{2}, meaning, for instance, the second homograph of the word *bow*. There are ways of keeping track of the order and the lexical entry to which that position refers, but a change in the original order of entries or the integration with other dictionaries in which the order differs would then require the update of all cross-references to any of the ordered entries. Since entries and senses are now identifiable throughout the data and graphs are not actually ordered, cross-references can be direct pointers to the entry or sense to which they refer.

The third advantage is intrinsically

15 https://www.w3.org/TR/rdf-sparql-query/

Asunción Gómez-Pérez is Vice-Rector for Research, Innovation and Doctoral Studies and Full Professor at Universidad Politécnica de Madrid. She is Head of the Department of Artificial Intelligence since 2008, Director of the Ontology Engineering Group since 1995, Academic Director of the Master's Degree in Artificial Intelligence since 2009, and Coordinator of the PhD Programme in Artificial Intelligence since 2009. Some of her main research areas are Ontological Engineering, Semantic Web, Linked Data, Multilingualism in Information and Management of Knowledge. She has been the coodinator of the OntoGrid, SemSorGrid4Env, SEALS, Interactivex and LIDER research projects, and she is currently taking part in three European H2020 research projects. She has also participated in numerous research projects of the Spanish National Plan of Basic Research, Networks, Special Actions and Technological Transference (ZENITH, Hundred, Advances, Profit, etc) and directed multiple projects of national and international enterprises. She is the head of the first node of the Open Data Institute in Spain and the main researcher of the first research project that uses IBM-Watson in a Spanish university (2015).
asun@fi.upm.es

The Ontology Engineering Group (OEG), led by Prof. Asunción Gómez-Pérez, is based at the Computer Science School at Polytechnic University of Madrid (UPM). It ranks eighth among the two hundred research groups of UPM and is widely recognized in Europe in the areas of Ontology Engineering, Semantic Infrastructure, Linked Data, and Data Integration. Its main research areas are Ontological Engineering, Open Science, Data-driven Language Technologies, Data on the Web, and Data Science. The OEG was the coordinator of LIDER, a European project that promoted the creation of a linked data-based ecosystem of interlinked multilingual language resources to support content analytics tasks. http://oeg-upm.net/

related with the first one, too: we can represent an "abstract" lexicon that gathers all the entries in a specific language. In other words, have a "pool" of lexical entries extracted from different dictionaries of the same or different type, monolingual or multilingual, without losing provenance information about which data comes from which dictionary. Thanks to an appropriate URI naming strategy, this pool of entries will grow dynamically (Gracia 2015) with each dictionary converted into RDF that has any information about an entry in that specific language.

If the approach mentioned above is applied in the conversion of multilingual dictionaries, for instance, Spanish-French and French-English, linking the French entries from the ES-FR dictionary with their corresponding entries in the FR-EN dictionary will bring us a fourth advantage: translation relations can be established through a language acting as a pivot (Villegas et al. 2016).

The fifth benefit concerns the onomasiological view that LD enables. The source dictionary has probably been compiled from a semasiological perspective, by putting the word as the center of attention and listing its different senses. Given the semantics by reference in LLD mentioned above, the synonym of a word and the word itself will point to the same concept, which is modeled as a node in the graph and has therefore a URI. Accessing that node will allow us to see which words lexicalize it, i.e. putting the concept as our focus and traversing the graph from it to the lexical elements related to it. This way of thinking is well illustrated in the case of multilingual dictionaries in LLD, where we can see how a concept is verbalized in different languages. The potential is however no less interesting in monolingual dictionaries. In the authors' work on the migration of language resources to LLD, definitions have been encoded at the level of the concept. Even though definitions can be fine-grained and are not presented in the form of keywords, SPARQL queries over them are feasible. For instance, we can search for concepts in whose definition the word *sunrise* occurs, which will yield the series of concepts that words like *dawn, morning, daylight*, etc. denote and which are semantically related, although these relations are not implemented. Through these concepts we could not only access *dawn, morning,* etc. but also their antonyms *dusk, twilight*, etc. Thus, by taking the concept as entry point we can get a set of concepts that are related but are not necessarily equivalent, which is not a trivial task when searching in a conventional online dictionary.

The sixth advantage is related to cross-references in the sense of any reference to another entry that might occur inside the lexicographic article: orthographical variants, synonyms, antonyms, genus terms, semantic types, etc. Not only are the entries reusable throughout the data (first advantage), but the pointers to them are now *typed* (Klimek and Brümmer 2015, McCracken 2015). This might not seem like an evident benefit to the user of online dictionaries, for whom the label *antonym* or a typographical mark may suffice, but typed properties allow users to perform queries not dependant on the (proprietary) format of the data and LD-aware systems to find any needed information. At the same time, by virtue of being defined in a public external vocabulary, e.g. LexInfo, the same properties can be reused in the conversion of other lexica of the same series into LLD, thus gaining interoperability. This responds to the need of standardization among the high number of heterogeneous annotation schemas, tagsets, and proprietary DTDs that are being used to create language resources.

Furthermore, given that these vocabularies are extensible, new properties and individuals or classes can be added. If the hierarchy defined in a linguistic ontology is not compatible with the view other domain experts might have, new vocabularies can be created and aligned to the ones already available. As we could experience during our work on the conversion of dictionaries to LD, a detailed comparison of the elements (and their classification) present in external vocabularies with the proprietary data model of a company specialized in lexicography is actually a significant step towards the improvement, refinement or even reconsideration of the elements that configure that data model.

As the last paragraphs suggest, the concept of *reusability* lies at the heart of LD. If the enterprise of compiling a dictionary is seen through the looking glass of LD from the very beginning, it will affect the whole process. Decisions such as, for example, keeping independent lexical entries for an entry and its homographs will have to be considered from the point of view of lexicography (two words that share form but are not related etymologically could thus be regarded as independent entries) and LD. At the same time, how do we model homographs in such a way that enables us to identify each entry but also to integrate content from another source that we do not know to which of the homograph entries it pertains? It will not be a matter of converting lexical data to LD, but of creating them from scratch in a reusable, interoperable and linguistically accurate way.

4. Benefits of a lexicon in linked data: micro-structure

The previous section dwelled on the benefits of representing a lexicon as LD but it did not deepen into the modeling of information present in a single lexicographic article. As opposed to lexical systems (Section 2), this information (definitions, grammatical data, syntactic frames, etc.) is also modeled as a graph.

Ideally, everything in the lexicographic entry can be modeled as a node (McCracken 2015) but, in general, and on the basis of *lemon-ontolex*, the representation revolves around lexical entries (`ontolex:LexicalEntry`), concepts, the relation between entries and concepts reified as lexical senses (`ontolex:LexicalSense`), word forms (`ontolex:Form`), definitions, phonetic representations, register, syntactic frames, etc. Relations between nodes have a well-defined domain and range, and, with actual data, every node will be an instance of a class defined in an ontology. Following the *lemon-ontolex* model, the English entry *cloud*, with the sample URI *lexiconEN/cloud-n* will have `rdf:type ontolex:LexicalEntry`, will denote as many `skos:Concepts` as senses or meanings it has, and the relation from the word to the concept will be encoded as `ontolex:LexicalSense`. Word forms (*cloud, clouds*) will be recorded at the `ontolex:Form` level, together with grammatical number information and phonetic transcription. Definitions, usage examples, etc. are likewise linked to the entry through edges and intermediate nodes.

On the one hand, one of the consequences of this configuration is that elements previously embedded in the lexicographic article become entry points in the graph and are no longer subsumed under any entry, since the hierarchy is lost. This implies that an idiom or collocation, for instance, will not be encapsulated under the container of the entry in which it was originally defined, but will be related to it with the suitable property. Since the idiom now becomes a node, we are able to link it to any other node from any other entry in the lexicon: *like a cat on a hot tin roof* could then be linked, for example, to the appropriate sense of *cat,* of *hot* and of *roof*, if desired, which will allow to access the idiom from any of those entries. Also, in the case of idioms and frequent collocations, we are creating new lexical entries that were not originally conceived as such in the lexicon. As lexical entries, they will be also linked to their corresponding `skos:Concept`(s), which brings us back to the possibility of an onomasiological perspective on the data.

On the other hand, thinking in terms of LD forces us to constantly question what is the nature of the relation between two pieces of information. An LD-native dictionary will require a specification on the part of lexicographers of which kind of relations between which type of elements will be encountered when modeling lexicographic articles. This brings us to the difference between compiling dictionaries with only the human as target, and creating them for (both humans and) computers. The fact that an XML tag, for instance, can occur at different levels in the dictionary entry (e.g. a geographical usage indication attached to a pronunciation vs. a geographical usage indication attached to a sense) seems straightforward enough for a human, but an NLP application needs to be able to distinguish between a description of a string (e.g. [kɑː] is the transcription of the British pronunciation of *car*) and the restriction on the *usage* of a sense (e.g. the *floor* with the meaning *the floor above the ground level floor* is only used in the UK). Modeling data as LLD thus entails a reflection of which information affects which elements, and which properties are the most suitable ones to be used in which case, taking all nuances and human implicit knowledge into account.

5. Conclusion and future lines of work

LLD emerge as a promising option to represent and publish current lexicographic projects and to serve as a structural backbone for undertaking new ones. They allow for the creation of an interoperable lexical network that is endowed with all the benefits that LD offers: data aggregation, easy discovery, LD-aware services compliance, improved data querying, sustainability and reusability. In this paper we have offered a brief overview of LLD, placing them in the context of lexical networks, and analyzing some of the benefits of the conversion of lexical data into LD in terms of macro- and microstructure. The modeling of lexicographic data to LLD poses challenges for which bridging the gap between LD experts and lexicographers is crucial. Moreover, the relation of LD to functional lexicography has not been explored to its full potential and, although there has been some work on RDF and OWL as building blocks for an architecture of mono- and plurifunctional dictionaries (Spohr 2011, 2012), this remains a challenging line of work, partly due to the increasing need of natural languages interfaces for the Web of Data. However, current trends in LD-based NLP and in publishing language resources as LD, including lexical data, show that we will be getting there hopefully soon.

Acknowledgments

This work is supported by the Spanish Ministry of Economy and Competitiveness through the project 4V (TIN2013-46238-C4-2-R), the Excellence Network ReTeLe (TIN2015-68955-REDT), the Juan de la Cierva program, and the Spanish Ministry of Education, Culture and Sports through the Formación del Profesorado Universitario (FPU) program. This contribution is inspired by the work towards the development of a linked data prototype for the Spanish dataset of the Global Series of K Dictionaries, carried out by the authors as part of the Linked Data Lexicography for High-End Language Technology Application (LDL4HELTA) project of Semantic Web Company and K Dictionaries.

LDL4HELTA

Linked Data Lexicography for High-End Language Technology Application (LDL4HELTA) is a 24-month EUREKA project (July 2015 – June 2017) within the framework of the Austria-Israel Bilateral R&D Agreement, carried out by Semantic Web Company (SWC, http://semantic-web.at/) and K Dictionaries (KD, http://kdictionaries.com/), with funding from the Austrian Research Promotion Agency (FFG) and the Israeli Office of the Chief Scientist (OCS). The aim is to combine multi-language lexical resources with semantic technologies expertise and develop new products and services for the international language technology market, in reply to the needs for language-independent, specific-language and cross-language solutions, to enable cross-lingual search and data management approaches. The main tasks consist of converting KD lexicographic data from XML to RDF, developing an API for enhanced data streaming and dissemination, and incorporating it in SWC's PoolParty Semantic Suite (https://poolparty.biz/). The RDF modeling is designed by the Ontology Engineering Group of Universidad Politécnica de Madrid (UPM), which is involved also in the word sense disambiguation aspects. An advisory board consists of Christian Chiarcos (Goethe University, Frankfurt), Orri Erling (Google), Asunción Gómez-Pérez (UPM), Sebastian Hellmann (Leipzig University), Alon Itai (Technion, Haifa), and Eveline Wandl-Vogt (Austrian Academy of Sciences). http://ldl4.com/

References

Bizer, C., T. Heath, and T. Berners-Lee. 2009. Linked data – the story so far. *International Journal on Semantic Web and Information Systems*, 5(3), 1-22.

Bosque-Gil, J., J. Gracia, E. Montiel-Ponsoda, and G. Aguado-de Cea. 2016. Modelling multilingual lexicographic resources for the Web of Data: The K Dictionaries case. In *Proceedings of GLOBALEX 2016 Workshop at the 10th Language Resources and Evaluation Conference (LREC 2016),* Portorož, (Slovenia).

Declerck, T., E. Wandl-Vogt, and K. Mörth. 2015. Towards Pan European Lexicography by Means of Linked (Open) Data. In *Electronic lexicography in the 21st century: Linking lexical data in the digital age. Proceedings of the eLex 2015 conference*, Herstmonceux Castle, UK, 342–355.

Fellbaum, C. (ed.). 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Fuertes-Olivera, P.A., and H. Bergenholtz. 2011. Introduction: The construction of Internet dictionaries. In Fuertes-Olivera, P.A., and H. Bergenholtz (eds.), *eLexicography: The Internet, Digital Initiatives and Lexicography*. London & New York: Continuum, 1-16.

Gader, N., V. Lux-Pogodalla, and A. Polguère. 2012. Hand-Crafting a Lexical Network With a Knowledge-Based Graph Editor. In *Third Workshop on Cognitive Aspects of the Lexicon (CogALex III)*, Mumbai, India, 109-125.

Gracia, J. 2015. Multilingual dictionaries and the Web of Data. *Kernerman Dictionary News*, (23), 1–4.

Klimek, B., and M. Brümmer. 2015. Enhancing lexicography with semantic language databases. *Kernerman Dictionary News,* (23), 5-10.

McCracken, J. 2015. The Exploitation of Dictionary Data and Metadata. In P. Durkin (ed.), *The Oxford Handbook of Lexicography*. Oxford: Oxford University Press, 501-514.

McCrae, J., G. Aguado-de-Cea, P. Buitelaar, P. Cimiano, T. Declerck, A. Gómez-Pérez, J. Gracia, L. Hollink, E. Montiel-Ponsoda, D. Spohr, and T. Wunner. 2012. Interchanging lexical resources on the Semantic Web. *Language Resources and Evaluation*, *46*(4), 701-719.

McCrae, J., C. Fellbaum, and P. Cimiano. 2014. Publishing and Linking WordNet using lemon and RDF. In *Proceedings of the 3rd Workshop on Linked Data in Linguistics*.

Mel'cuk, I. 1996. Lexical functions: a tool for the description of lexical relations in a lexicon. *Lexical functions in lexicography and natural language processing*, (31), 37-102.

Miller, G.A. 1995. WordNet: A Lexical Database for English. *Communications of the ACM 38*(11), 39-41.

Montiel-Ponsoda, E., G. Aguado de Cea, A. Gómez-Pérez, and W. Peters. 2008. Modelling Multilinguality in Ontologies. In *The 22nd International Conference on Computational Linguistics (COLING 2008)*, August 18-22, Manchester, UK, 67-70.

Polguère, A. 2012. Like a lexicographer weaving her lexical network. In *Proceedings of CogALex-III Workshop of the 24th International Conference on Computational Linguistics (COLING 2012)*, December 8-15, Bombay, India. 1-4.

Polguère, A. 2014. From Writing Dictionaries to Weaving Lexical Networks. *International Journal of Lexicography*, *27*(4), 396-418.

Spohr, D. 2011. A Multi-layer Architecture for "Pluri-monofunctional" Dictionaries. In Fuertes-Olivera, P.A., and H. Bergenholtz (eds.), *eLexicography: The Internet, Digital Initiatives and Lexicography*. London & New York: Continuum, 103-120.

Spohr, D. 2012. *Towards a Multifunctional Lexical Resource: Design and Implementation of a Graph-based Lexicon Model*. Lexicographica Series Maior (141). Berlin: de Gruyter.

Swanepoel, P.H. 1994. Problems, theories and methodologies in current lexicographic semantic research. In W. Martin et al. (eds.), *Proceedings of the Sixth International Euralex Congress,* Amsterdam, 11-26.

Villegas, M., M. Melero, N. Bel., and J. Gracia. 2016. Leveraging RDF graphs for crossing multiple bilingual dictionaries. In *Proceedings of the 10th Language Resources and Evaluation Conference (LREC 2016),* Portorož, (Slovenia).

Wandl-Vogt, E. 2015. How to innovate lexicography by means of research infrastructures: The European examples of DARIAH, CLARIN and COST IS 1305 ENeL [slides]. http://www.slideshare.net/ewv/how-to-innovate-lexicography-by-means-of-research-infrastructures/ (June 5, 2016).

From dictionaries to cross-lingual lexical resources

Guadalupe Aguado-de-Cea, Elena Montiel-Ponsoda, Ilan Kernerman and Noam Ordan

1 Introduction

While the number of general resources that are connected as part of the linked open data paradigm increases, the need to relate and link linguistic data in multiple languages as a result of this trend has rocketed as well. The vision of a universe that allows linguistic information from different resources to be interlinked has attracted many scholars in search of "the magic wand" for solving the everlasting problem of the Tower of Babel, which now includes languages for machines in addition to human users. Currently, most linguistic resources are still in proprietary formats, making it difficult to be linked and interoperate on the Web. To achieve that envisioned linked cloud of linguistic resources, several issues have to be addressed, from representation models to linking processes, from querying interfaces to dataset maintenance solutions.

Great advances in methodologies and techniques for the publication of linked data are laying solid foundations for turning independent databases into a boundless cloud where users can make queries in an integrated environment using dedicated, standardized querying languages, thus catering for interoperability as well as fostering univocity of the elements described. Linked data relies on the Resource Description Framework (RDF) data model[1] as the main mechanism applied to describe data. These data in turn are linked to other similarly modelled data, and ultimately retrieved and manipulated by using Web standards such as the SPARQL[2] query language.

Many language resources have seen the advantages of complying to this new paradigm, and are currently available as part of the Linguistic Linked Open Data (LLOD) Cloud[3], a sub-cloud of the linked open data cloud that brings together linguistic resources formalized in RDF (from lexicons, dictionaries, and terminologies to metadata repositories and corpora). However, as in the case of the traditional Web, the LLOD is mainly English-oriented, though more non-English data sources are increasingly being published. As stated by Gracia et al. (2011), the new challenge is to overcome language barriers if we aim to attain a truly multilingual Semantic Web.

WordNet[4] (Fellbaum 1988), for example, which is the most widely used lexico-semantic resource in English with more than 117,000 synsets (sets of synonyms that account for a concept), has recently undertaken a new role in constructing the Semantic Web (Berners Lee et al. 2001). The W3C draft RDF/OWL Representation of WordNet[5] has defined URIs for the synsets covered by the WordNet lexical database. Many other efforts have been devoted to link WordNet to other resources. McCrae et al. (2012) used WordNet together with Wiktionary as a case study of the possible transformation of lexical resources into linked data compatible formats. In McCrae et al. (2014), the authors provide RDF-compliant Wordnet with links to other lexical resources, such as VerbNet[6], Lexvo[7] or lemonUby.[8]

As for multilingual linguistic resources which are part of the current LLOD cloud, it is worth mentioning IATE RDF[9] (Cimiano et al. 2015), AGROVOC[10] and EUROVOC in SKOS[11], or the APERTIUM[12] series of bilingual dictionaries (all of which are navigable and searchable from Datahub[13]). Several chapters of DBpedia[14] are now available in different languages, as well as some language versions of EuroWordNet (the Basque[15] and Catalan[16] versions present a case in point). However, what still remains

1 https://www.w3.org/standards/techs/rdf#w3c_all/
2 http://www.linkeddatatools.com/querying-semantic-data/
3 http://linguistic-lod.org/llod-cloud/
4 https://wordnet.princeton.edu/
5 https://www.w3.org/TR/wordnet-rdf/
6 http://verbs.colorado.edu/~mpalmer/projects/verbnet.html/
7 http://www.lexvo.org/
8 http://lemon-model.net/lexica/uby/
9 https://datahub.io/es/dataset/iate-rdf/
10 https://datahub.io/es/dataset/agrovoc-skos/
11 https://datahub.io/es/dataset/eurovoc-in-skos/
12 https://datahub.io/es/dataset/apertium-rdf/
13 https://datahub.io/
14 http://linghub.lider-project.eu/datahub/dbpedia/
15 http://linghub.lider-project.eu/datahub/basque-eurowordnet-lemon-lexicon-3-0/
16 http://linghub.lider-project.eu/datahub/catalan-eurowordnet-lemon-lexicon-3-0/

Guadalupe Aguado-de-Cea is Professor at Universidad Politécnica de Madrid (UPM). She received both her MSc in Translation and PhD in English Philology from Universidad Complutense de Madrid, and has been a member of the Ontology Engineering Group at UPM since 1996. Her current research activities include, among others: terminology and ontologies, the representation of lexical knowledge in ontologies, multilinguality in linked data, specialized languages as well as the linking between the ontological field and the natural language field, especially in its application to the Semantic Web. She has participated in several standardization projects, such as the Ontology Lexica Community Group (Ontolex) in the W3C, in particular regarding the representation of translation relations among languages with a view on the multilingual Web. She is the President of the Spanish Association for Terminology, and convenor of the AENOR CTN_191 Terminology Committee, the corresponding Spanish Committee of ISO TC 37.
lupe@fi.upm.es

Elena Montiel-Ponsoda is Associate Professor at the Applied Linguistics Department at Universidad Politécnica de Madrid (UPM) since 2012, and member of the Ontology Engineering Group since 2006. She got her PhD on Applied Linguistics from UPM in 2011. Her research interests are at the intersection between translation (and terminology) and knowledge representation, including among others: ontology localization and lexicalization, lexico-syntactic patterns for ontology development, functional models for deep semantics analysis, sentiment analysis, and linguistic linked data for content analytics. She is currently working on the representation of lexical resources according to the linked data paradigm, specifically, on how translation relations can help in the construction of the multilingual Web of Data.
emontiel@fi.upm.es

a challenging issue is the flawless linking of complementary resources in different natural languages. By complementary resources, we refer to resources that deal with the same (or closely related) parcels of knowledge, be it general or domain-specific knowledge, whose metadata descriptions as well as actual data are in different natural languages. In this sense, we argue that Semantic Web approaches and technologies are ripe enough to offer viable solutions to the linking issue in a principled manner.

Our objective in this contribution is to report on our experience in modelling the linked data version of the Spanish set of the K Dictionaries (KD) multi-language Global Series that will serve to transform a multilingual dictionary into a cross-lingual lexical resource. We would like this to set ground for discussion to define open issues for the linkage of lexical data in multiple languages, and some solutions are suggested on the base of *de-facto* standard *lemon-ontolex* model[17], initially designed to serve as an interface between an ontology and the natural language descriptions that lexicalize the knowledge represented in it, and currently widely adopted for exposing linguistic resources as linked data. Specifically, we describe how multilingual information in the RDF version of KD's dataset has been represented according to the *vartrans* module, a *lemon-ontolex* module for representing translations and term variants, and how this could contribute to enhance interoperability among the different language versions of the Global Series.

The paper is further structured as follows. In the next section we refer to the background and motivation, i.e. approaches to linking multilingual lexical and/or conceptual resources. Section 3 introduces the KD approach and Section 4 presents the formal solution we have adopted for its Spanish dataset in the linked data model, specifically, the *lemon-ontolex vartrans* module. The actual modeling of the Spanish dataset from the XML proprietary format of the dictionaries is spelled out in Section 5. In Section 6 we list some advantages of complying to this or similar formalisms in the context of the linked data paradigm, and our conclusions are presented in Section 7.

2 Background and motivation

When approaching this issue in the Semantic Web field, it is inevitable to refer to a former, much older discussion on how to bring together lexicons in different languages and how to solve discrepancies resulting from the idiosyncratic categorization of each language system/culture, some of which are reflected in the way different linguistic features, such as gender, pronouns or classifiers, are encoded (cf. Fellbaum and Vossen 2007). One of the resources that better materializes (and tries to solve) this problem is EuroWordNet (Vossen 1998), and subsequently derived projects such as MultiWordNet (Pianta et al. 2002). Broadly speaking, such databases connect wordnets or lexicons in different languages via a set core of categories, the so-called Interlingual Index, based on Princeton WordNet (Miller 1995). In the case of EuroWordNet there is an implicit bias towards English synonym sets which allegedly stand for concepts realized lexically by lexical items in different languages, and, in the case of MultiWordNet, the bias is more explicit, because the English WordNet is literally translated into the various languages, and gaps are declared by free translations that stand for those concepts, allowing linked concepts/synsets to percolate through the gaps.

In this regard, we would argue that different language-culture couplings (we see this as a binomial) can exhibit different levels of granularity when representing and categorizing knowledge. Even among culturally-related languages, such as Italian and English, it has been shown that a medium-sized dictionary of English to Italian contains around 7.8% lexical gaps, where there is no equivalence and a free translation is needed to fill the gap (Bentivogli and Pianta 2000). Therefore, and in order to address these issues, the Global WordNet Grid (Fellbaum and Vossen 2007; Vossen et al. 2016) initiative aims at providing a platform for centralising all wordnets and their linkage, and coordinating the inclusion of new concepts for multiple languages. As such, this latter approach represents an important step towards a more principled solution to the multilingual (still unresolved) issue.[18]

Another approach that also builds on WordNet, but which has been born in the Semantic Web era, is BabelNet (Navigli and Ponzeto 2012). This is a semantic network and ontology that aims at bringing together words and terms in different languages, from various resources, which refer to the same concept, with the objective of serving as valuable sources of translation or equivalent relations. According to Moro and Navigli (2015), in BabelNet it is possible "to find the

17 https://www.w3.org/community/ontolex/wiki/Final_Model_Specification/

18 cf. http://compling.hss.ntu.edu.sg/omw/

concept *medicine* (bn:00054128n), which is represented by both the second word sense of *medicine* in WordNet and the Wikipedia page *Pharmaceutical drug*, among others, together with synonyms such as *drug* and *medication* in English and lexicalizations in other languages, such as *farmaco* in Italian and *medicamento* in Spanish". In this way, BabelNet combines the general-specific approach taken from WordNet with the specific knowledge extracted from Wikipedia (and other resources, e.g. OmegaWiki). As for the English-language bias issue, it is probably propagated to this resource, since WordNet is taken as a starting point. However, it can also be reduced, because of the use of Wikipedia entry pages for categories not initially included in the original WordNet.

Apart from acknowledging the great value of such a resource, we have also spotted some flaws that will undoubtedly be solved in the future, and which are probably due to automating the linking process. For instance, some synsets contain words that belong to different categories. An example is the synset for *paella* (typical Spanish rice dish), which also includes the pan used to cook it. As for the translations in BabelNet, when different options are offered, we would suggest that additional information is required, such as confidence scores associated to the proposed translation, pragmatic restrictions (for instance, the frequency with which a word in language A is translated with the proposed equivalent in language B), or directionality of the translations. Means such as these would positively contribute to enhance this resource's functionality.

All in all, and although many advances have been made in the alignment and linking of resources in different languages, it is still necessary to cater for certain aspects in order to make the most of the multilingual information contained in such resources.

3 The K Dictionaries approach

The dictionary data used as input in this research belong to the Global Series of K Dictionaries (KD)[19]. KD is a technology-oriented-content creator that specializes in developing pedagogical and multilingual lexicographic data. In 2005 it launched the Global Series, which today includes lexical resources for 24 languages. The approach followed in this series is to compile for each language a core vocabulary as a standalone project, and have it translated to other languages in more projects. In other words, there is no bias towards any language, each is represented on its own terms, and only at a later phase it is translated to another, creating a pair-specific, and thus pair-sensitive, interlingual representation.

The outset of each language dataset in this series concerns mapping its components to identify, categorize and interlink them, including semantic and grammatical information. Each language core then serves as a base for adding translation equivalents in other languages and developing bilingual and multilingual versions. All the different language datasets share the same common methodological framework and technical infrastructure. The entries in the different languages also have the same microstructure, which still enables each one to convey its peculiarities. The data is structured in XML format and is currently being modeled in RDF. The French dataset, for instance, has the most extensive multilingual reach so far with 18 language pairs, the German lexical dataset groups 8 more languages, Spanish has 7, Japanese – 7, English – 6, Norwegian – 6, etc. Now that several language sets have become so lexically rich, they are ripe to start networking with each other, such as by connecting L2 translations to their corresponding entries in the L1 lexical dataset and from there on to translations in other languages, and so on.

As explained in the introduction, we reflect here on some interesting issues spotted when transforming the Spanish lexical core of the Global dataset, focusing on multilingual ones. We leave aside the methodology followed in the modeling part, which has been described in greater detail in Bosque-Gil et al. (2016a and 2016b), and move on to the resulting representation of translations in the proposed model.

4 *lemon-ontolex* at a glance: The *vartrans* module

In order to link and represent the linguistic data included in KD's Global Spanish dataset we relied on the *lemon-ontolex vartrans* module. It presents wide possibilities to link lexical senses and variants in different languages from the same or different data sets. As shown in Figure 1, the lexico-semantic generic class addresses the relation between two lexical entries or two lexical senses. This relation is established by means of two properties: `lexicalRel` and `senseRel`. Thus, `lexicalRel` relates two lexical entries that are grammatically or stylistically connected, such as acronyms, derivatives and other forms.

The second class, `senseRel`, represents the relation between two senses whose meanings are related. Not only can

19 http://kdictionaries.com/

Ilan Kernerman is CEO of K Dictionaries, leading its lexicographic development and international cooperation. He edits and publishes *Kernerman Dictionary News*, co-edited and published two collections of conference papers (1998, with Tom McArthur, and 2010, with Paul Bogaards), and is associate editor of *Lexicography – Journal of Asialex* and guest co-editor of the special *IJL* issue on bilingual learners' dictionaries (2016, with Arleta Adamska-Sałaciak). His interests include multilingual and pedagogical lexicography, and interoperability with NLP and knowledge systems. Currently he is president of Asialex (2015-2017) and on the preparatory board of Globalex.
ilan@kdictionaries.com

Noam Ordan studied translation, linguistically and computationally, and completed his PhD at Bar Ilan University under the supervision of the late Miriam Schlesinger. He has published extensively, in particular on automatically identifying translated texts and statistical machine translation, worked as researcher and teacher in universities in Israel and Germany, and took part in various projects in the industry. Currently he coordinates research innovation at K Dictionaries and designs algorithms for using human-crafted lexicographic data for computational tasks, such as cross-lingual information retrieval. Dr Ordan also serves as an adjunct teacher at the English Language Department at the Arabic Academic College in Haifa.
noam@kdictionaries.com

lexico-semantic relations, such as synonymy, antonymy or hypernymy-hyponymy be represented in this way, but also term variants and translations. The purpose of such a representation is to account for two lexical senses of terms (in the same or different language) that are semantically related in the sense that they can be exchanged in most contexts, but their surface forms are not directly related. Additionally, other types of semantic and pragmatic information, such as dialectal, registerial, chronological, discursive, and dimensional variation can also be captured by `senseRel`.

5 Modelling multilingual entries in the KD data with *vartrans*

The starting point in the transformation of the multilingual information (translations) contained in KD's Global Spanish dataset was a 'Translation cluster' that encompassed a set of translations for the original Spanish lexical entry, including syntactic-semantic and pragmatic information about the translations (e.g. grammatical gender), and usage examples of the headword (commonly a short phrase), as well as translations of those examples.

See Example 1 for the XML encoding of the headword *acalorado* (*heated*), which contains a synonym, namely, *agitado* (*lively* or *passionate*), a definition, *que es muy animado* (*of a discussion or debate, that is heated*), and translations into Dutch (*verhit* and *vurig*) and Norwegian (*ivrig*, *oppsatt*, and *opphetet*). Moreover, this sense of *acalorado* is complemented with a usage example (*una sesión acalorada*), and its equivalents in Dutch (*vurige zitting*) and Norwegian (*et opphetet møte*), respectively, are all included in the ExampleCtn type and identified by means of a translation cluster identifier given in the XML, TC00001664.

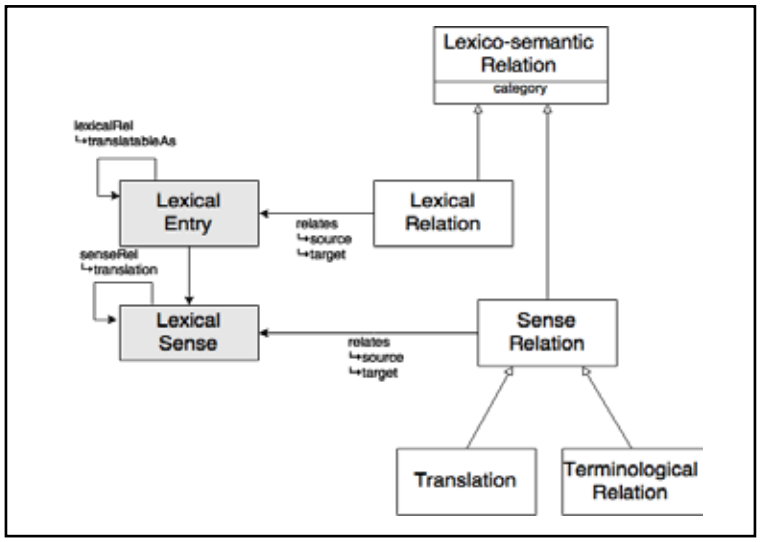


Figure 1. Classes and properties in the *vartrans* module

According to *lemon-ontolex*, a dictionary entry or headword in the KD set is modeled as an `ontolex:LexicalEntry` and its corresponding `ontolex:LexicalSense` and `skos:Concept`, as can be seen in Figure 2. Then, according to the *vartrans* module, synonym relations are modeled as relations between lexical senses that point to (`ontolex:reference`) the same concept (`skos:Concept`). Thus, for example, the lexical entry for the headword *acalorado* is linked to its corresponding sense and concept, and an artificial sense is created for the synonymous lexical entry *agitado*, so that a sense relation of the type synonymy can be established between them. Should *agitado* have also its own headword in the dictionary, a link could be established between the lexical senses later on, or lexical senses could be merged. Both lexical senses refer to the same `skos:Concept`, and a definition is also attached to the latter.

Similarly, translations are modeled as relations among lexical senses. Again, if we analyze Figure 2, the lexical sense for the entry in the source language (*acalorado*) is available, and the sense for the target language (*verhit*) has to be artificially created, since no pointer to that entry in other dictionaries is provided in the XML data (once the Dutch and Norwegian datasets are converted to RDF, these entities can support the automatic linking and growth of both datasets). The usage examples that accompany the senses are represented by means of the property `skos:example` and the class `kd:UsageExample`. Moreover, examples of usage are commonly translated into other languages and grouped by the `kd:TranslationExampleCluster`, a grouping made in the original datasets and maintained here.

The modeling solution proposed by the *vartrans* module for representing a translation relation by means of a reified class instead of a property or relation facilitates the further description of the translation object. In this sense, `translationSource` and `translationTarget` can be further specified, as done for the current version of the KD Spanish set. Also, other features that describe a certain translation relation could be added. For example, a confidence value can be assigned to the translation pair if available. A context could be determined to restrict the validity of the translation pair and differentiate it from other possible translations of the original entry into the target language. In fact, if we consider the usage examples available for *acalorado* in the XML dataset, *una sesión acalorada* (*a heated session*) has been translated into Dutch as *vurige zitting*, and not as *verhite zitting*, which was the synonym provided.

And the same happens with the Norwegian alternatives, the phrase is translated as *et opphetet møte*, and as learners of Norwegian we may wonder if the other two synonyms offered for *opphetet*, namely, *ivrig* and *oppsatt*, can be interchangeably used in that phrase.

Additionally, we may want to specify the type of translation relation that exists between a pair of translation equivalents. Gracia et al. (2014) propose a classification of translation equivalents into three types: direct equivalents (lexical entries in the translation pair that are semantically equivalent), cultural equivalents (lexical entries that are not semantically equivalent, but are pragmatically so), and lexical equivalents (the target lexical entry – or translation equivalent – verbalizes the original entry in the target language but is not a semantic or pragmatic equivalent). For more details we address the interested reader to the above-cited paper.

Therefore, apart from specifying the origin and target of the translation pair, the other descriptions that could further enrich the information related to were not available in the original source and have not been implemented in the current version. That does not mean such descriptions could not be added or imported from another resource that contains data to that respect. In fact, this is one of the main benefits of adopting the linked data paradigm, namely, being able to link to resources containing complementary information.

```
<SenseGrp identifier="SE00000730" version="1">
        <Synonym>agitado</Synonym>
        <Definition>que es muy animado</Definition>
       <TranslationCluster identifier="TC00001663" text="que es muy
animado" type="def">
         <Locale lang="nl">
          <TranslationBlock>
           <TranslationCtn>
            <Translation>verhit</Translation>
           </TranslationCtn>
           <TranslationCtn>
            <Translation>vurig</Translation>
           </TranslationCtn>
          </TranslationBlock>
         </Locale>
         <Locale lang="no">
          <TranslationBlock>
           <TranslationCtn>
            <Translation>ivrig, oppsatt, opphetet</Translation>
           </TranslationCtn>
          </TranslationBlock>
         […]
       </TranslationCluster>
        <ExampleCtn type="sid" version="1">
         <Example>sesión acalorada</Example>
          <TranslationCluster identifier="TC00001664" text="sesión
acalorada" type="exmp">
          <Locale lang="nl">
           <TranslationBlock>
            <TranslationCtn>
             <Translation>vurige zitting</Translation>
            </TranslationCtn>
           </TranslationBlock>
          </Locale>
          <Locale lang="no">
           <TranslationBlock>
            <TranslationCtn>
             <Translation>et opphetet møte</Translation>
            </TranslationCtn>
           </TranslationBlock>
          […]
         </TranslationCluster>
```

Example 1: XML with the translations in Dutch and Norwegian of the Spanish headword *acalorado* sense of *heated*

6 Advantages of cross-lingual lexical resources

Our reflections in this paper are made to point out some advantages of linking multilingual datasets in the aim of getting the most of the multilingual data value chains in the cloud of linked data. We argue that the linked data representation formalism offers an innovative way of bringing together resources in which either the vocabularies or models, or the data itself, are described in different natural languages, contributing to the construction of a truly multilingual Semantic Web. The challenge here is to account for as comprehensibe as possible specifics of each language taken *individually* while at the same time to represent links with meaningful labels across languages within a multilingual graph.

In the specific case of the lexical resources under examination, we argue that by representing translations as links between lexical senses (and, in turn, lexical entries), whenever new datasets that contain information in the target languages are also represented according to this paradigm, links will be flawlessly established. As already mentioned in previous sections of this paper, once the different datasets of the Global Series are available in RDF, links will be established among the different entities, contributing to an automatic growth of the resources. If we take the example of KD's Global Spanish dataset, since it contains translations into Brazilian Portuguese, Dutch, English, Japanese, and Norwegian it is reasonable to assume that relying on those translations, links will be easily created among the different datasets.

Although this is still a visionary

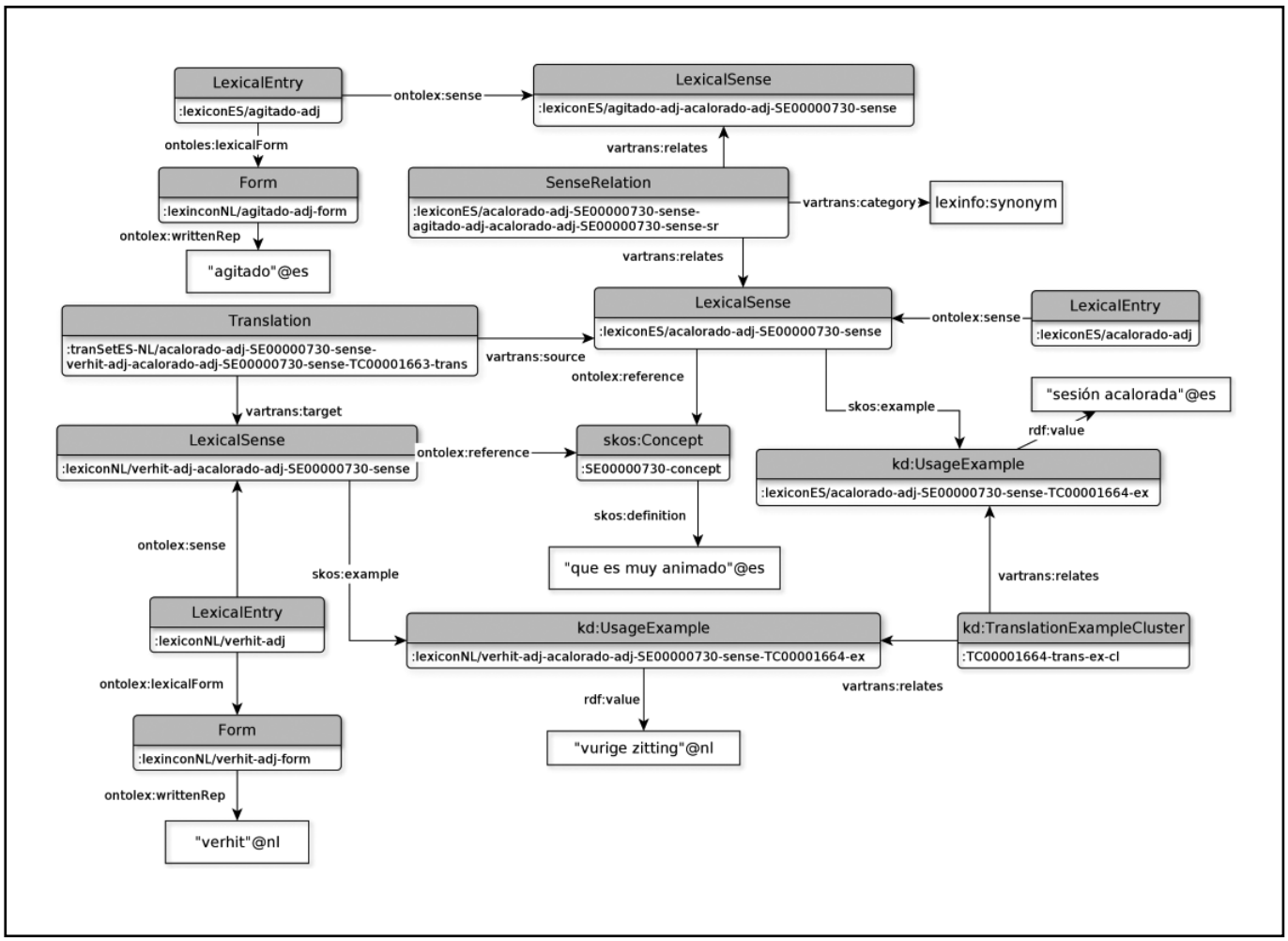


Figure 2. Modeling of a KD multilingual entry with *lemon-ontolex*

concept, representing lexical resources according to this approach will enable the emergence of a cross-lingual graph in a bottom-up fashion. This will maintain the distributed fashion of the linked data graph, and datasets will be easily connected, disconnected or contextualized for specific users and uses.

Contrary to the approaches described in state-of-the-art projects within the Global Grid initiative, we believe no common set of concepts or intermediary conceptualization would be needed to establish cross-lingual relations, but links would emerge among datasets at a different pace. Put differently, instead of relying on a common conceptualization to act as intermediary, the *burden of the cross-lingual connection* would be carried by the links.

At a monolingual level, since the relation between synonyms or terminological variants has been also reified in the `TerminologicalRelation` class, we could also determine precisely if a certain synonym or term is used in a specific context, or if all the synonyms related to the same concept can be interchangeably used. In the example of the BabelNet *medicine* concept mentioned in Section 2, we could identify accurately the specific uses of *medicine* versus *Pharmaceutical drug*, *drug* or *medication*. Are they used in the same contexts? Which is the most appropriate translation for *medicamento* in Spanish in an informal setting?

This is also specifically relevant in those cases in which complex linguistic descriptions are associated to conceptual structures. Let us consider the example of *biosanitary waste,* in general, and *hospital waste*, only for the waste produced in hospitals. If the difference between these concepts is established at the conceptual level, the two terms will most probably be associated to two different concepts. Conversely, if only one concept is represented in the ontology, we may still want to account for both terminological variants in the linguistic model, and explicitly state the motivation behind each denomination. In this way, we would also facilitate the linking of this data source to another data source contained in a different dataset and to which only the term *biosanitary waste* has been associated.

7 Conclusions

Following the experiences in this project we can claim that the publication of lexical and terminological resources as linked data will result in an enriched unified graph of lexical entries, senses and translations on the Web. Consequently, more information (additional notes, glosses, descriptions) will be retrieved by querying the linked data resources by means of SPARQL queries. Moreover, they could be enriched with pictures, audio, and the like, as has been successfully implemented in BabelNet, for example. However, having stated the benefits of linking linguistic resources, and more specifically the advantages of this initiative when applied to multilingual lexical resources, we are also aware of the challenges that still need to be tackled and that have been discussed in Section 6.

Acknowledgements

This work is supported by the 4V Spanish National Project (TIN2013-46238-C4-2-R), the Spanish Excellence Network ReTeLe (TIN2015-68955-REDT), and the LDL4HELTA project under the EUREKA program.

References

Bentivogli, L., and Pianta, E. 2000. Looking for lexical gaps. In *Proceedings of the Ninth EURALEX International Congress, EURALEX 2000*, pp. 663-669.

Berners-Lee, T., Hendler, J. and Lassila, O. 2001. The Semantic Web. *Scientific American*, May 2001. 29-37.

Bosque-Gil, J., Montiel-Ponsoda, E. Gracia, J. and Aguado-De-Cea, G. 2016a. Terminoteca RDF: a Gathering Point for Multilingual Terminologies in Spain. *12th International Conference on Terminology and Knowledge Engineering (TKE 2016*).

Bosque-Gil,J.,Gracia,J.,Montiel-Ponsoda, E. and Aguado-De-Cea, G. 2016b. Modelling multilingual lexicographic resources for the Web of Data: The K Dictionaries case. In *Globalex 2016 Lexicographic Resources for Human Language Technology.*

Cimiano, P., McCrae, J., Rodríguez-Doncel, V. Gornostay, T. Gómez-Pérez and B. Simoneit. 2015. Linked Terminology: Applying Linked Data Principles to Terminological Resources. In *Proceedings of the 4th Biennial Conference on Electronic Lexicography*.

Fellbaum, C. (ed.). 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Fellbaum C. and P. Vossen 2007. Connecting the Universal to the Specific: Towards the Global Grid. In *Proceedings of The First International Workshop on Intercultural Collaboration (IWIC 2007)*, Kyoto, Japan, January 25-26.

Gracia, J. Montiel-Ponsoda, E. Cimiano, P. Gómez-Pérez, A. Buitelaar, P. and J. McCrae. 2011. Challenges for the multilingual Web of Data. In *Web Semantics: Science, Services and Agents on the World Wide Web* 11: 63-71.

Gracia, J. 2015. Multilingual dictionaries and the Web of Data. *Kernerman Dictionary News*, (23), 1-4.

McCrae, J., Cimiano, P., and Montiel-Ponsoda, E. 2012. Integrating WordNet and Wiktionary with lemon. In C. Chiarcos, S. Nordhoff, S. Hellmann, (eds.), *Linked Data and Linguistics: Representing and Connecting Language data and Language Metadata. Heidelberg & New York: Springe*r, 25-34.

McCrae, J., Fellbaum, Ch. and Cimiano, P. 2014. Publishing and Linking WordNet using lemon and RDF. In *Proceedings of the 3rd Workshop on Linked Data in Linguistics*.

Miller, G. A. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11), 39-41.

Montiel-Ponsoda, E., Bosque-Gil, J.,Gracia, J. Aguado-de-Cea, G. and Vila-Suero, D. 2015. Towards the integration of multilingual terminologies: an example of a linked data prototype. *TIA 2015, Granada, Spain*.

Moro A. and Navigli R. 2015. SemEval-2015 Task 13: Multilingual All-Words Sense Disambiguation and Entity Linking. In *Proceedings SemEval 2015*, Denver, Colorado, June 4-5, 2005, 288-297.

Navigli, R. and Ponzetto, S. P. 2012. BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. *Artificial Intelligence*, 193, 217-250

Pianta, E., Bentivogli, L., and Girardi, C. 2002. Developing an aligned multilingual database. In *Proceedings of the 1st International Conference on Global WordNet*.

Vossen, P. 1998. *A multilingual database with lexical semantic networks*. Dordrecht: Kluwer Academic Publishers.

Vossen, P., Bond, F. and McCrae, J. 2016. Toward a truly multilingual GlobalWordnet Grid. In *Proceedings of the 8th Global Wordnet Conference* (GWC 2016).

This paper was presented at META FORUM 2016 in Lisbon, Portugal on 5 July 2016. http://meta-net.eu/events/meta-forum-2016/

Elena Montiel-Ponsoda at META

Interns @ KD 2015-2016

- Universitat Jaume I, Castelló

Lucia Belles Calvera.
Lidia Gallen Martinez.
Miriam Martinez Garcia.
Catalan-Spanish index and *English-Catalan dictionary*

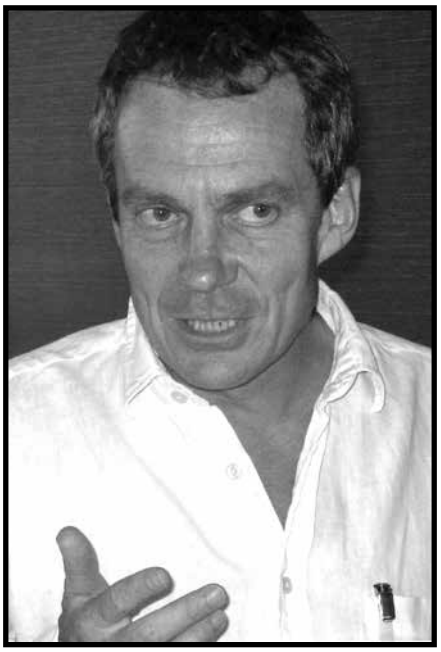
- KU Leuven, Antwerp

Rhiannon Telery Hincks.
English-Welsh dictionary
Zrinka Knezovic.
Croatian-English index
Liubava Panchenko.
Ukrainian-English index

- Université de Lorraine, Nancy

Pauline Pierrot.
French dictionary

Adam Kilgarriff Prize



Adam Kilgarriff at Euralex 2010

At last year's eLex conference in Herstmonceux Castle (UK), almost every paper and poster included at least one reference to Adam Kilgarriff's enormous body of work – a vivid demonstration (if any were needed) of Adam's extraordinary impact on the fields he had worked in. A group of us met there to discuss setting up a prize in honour of our dear friend and gifted colleague, who died in May 2015. We are now pleased to announce the launch of the Adam Kilgarriff Prize, which will be awarded every two years, in conjunction with the eLex conference series. The Prize is aimed at younger researchers and is intended to recognise outstanding work in any of the fields which Adam enriched with his remarkable intellect and original thinking.

Almost uniquely, Adam was a major figure in three quite distinct communities: natural language processing (NLP), lexicography, and corpus linguistics. He was an enthusiastic, insightful, and prolific contributor to each of these fields, but perhaps his best work straddled all three, and few people have had such a profound impact on the practice of contemporary lexicography in particular. Through numerous collaborations with dictionary makers, Adam brought to bear his NLP skills and can-do approach to provide elegant solutions to many of the challenges which lexicographers face day to day. Issues such as word sense disambiguation, corpus building, and headword-list development all engaged Adam's attention – and lexicography is the richer for his interventions. In many cases, he proposed a software solution, and this led to the development of tools such as the GDEX (good example) algorithm, now widely used (in several languages) as a computational shortcut for the process of finding in a corpus the most appropriate example sentences and phrases for a dictionary.

During a research project at Brighton University in the late 1990s, Adam conceived – with his co-researcher David Tugwell – the notion of a Word Sketch. This would provide a one-page overview of a word's most typical behaviour, summarizing the most frequent and significant ways in which it would combine with other words in text. An experimental version was used during the development of *Macmillan English Dictionary for Advanced Learners* (2002), and before long Word Sketches had become an essential resource in the lexicographer's toolbox. Harnessing Word Sketch technology to a powerful concordancer led to the birth of the Sketch Engine. Under Adam's leadership, this suite of corpus-analysis tools was continuously improved and enhanced, to become an industry-standard package for dictionary publishers as well as for other linguistic undertakings worldwide.

There is much more, and this short account can hardly do justice to Adam's amazing achievements. It is hard to believe that one individual could have done so much in such a short lifetime, and we hope that the Adam Kilgarriff Prize will be a fitting memorial to Adam's life and work.

Details of the Prize – and how to apply for it – can be found at: http://kilgarriff.co.uk/prize/.

Michael Rundell
Chair of Trustees, Adam Kilgarriff Prize