Increasing information accessibility on the Web: a rating system for specialized dictionaries

Valeria Caruso*, Anna De Meo*, Vincenzo Norman Vitale°
*Università degli Studi di Napoli 'L'Orientale',
°Università degli Studi di Napoli Federico II
vcaruso@unior.it, ademeo@unior.it,
vincenzon.vitale@studenti.unina.it

Abstract

English. The paper illustrates the features of the *WLR* (Web Linguistic Resources) portal, which collects specialized online dictionaries and asses their suitability for different functions using a specifically designed rating system. The contribution aims to demonstrate how the existing tool has improved the usefulness of lexicographical portals and how its effectiveness can be further increased by transforming the portal into a collaborative resource.

Italiano. *Questo contributo descrive le caratteristiche del portale WLR (Web Linguistic Resources) che raccoglie dizionari specialistici della Rete e ne stima l'utilizzabilità per diverse funzioni, avvalendosi di uno specifico sistema di valutazione. Viene quindi mostrato come questo strumento incrementi l'utilizzabilità dei portali lessicografici finora sviluppati e come la sua efficacia possa essere ulteriormente migliorata trasformandolo in risorsa collaborativa.*

1 Introduction

This paper sketches out the current features and an upcoming new application of a rating system designed to assess online specialized dictionaries. The system evaluative parameters are managed through a relational database accessible for free online at the *Web Linguistic Resources* (*WLR*) site. These parameters are used to identify the best available dictionaries to satisfy different types of information needs experienced by the Internet surfers, while the assessment procedure has been designed to be flexible and can be readapted to estimate the supportive value of other resources as well, like grammars or corpora. On the other side, once the score assignment for each dictionary feature has been decided, grades are given automatically by the database.
The assessment procedure is straight and strictly operationalized (Swanepoel, 2008, 2013), and it can be used as a guided process to collect data provided by the users themselves. The system is in fact going to be updated and transformed in a collaborative (Carr, 1977) dictionary portal, collecting forms that have been filled in by the Web surfers themselves.

2 Information overload on the Internet

The *WLR* dictionary portal has been designed as a tool that can offer assistance to solve different problems concerning specialized knowledge and lexicon that Web users might experience on different occasions in their lives. For example, if they need to understand specific concepts belonging to some technical fields, like a journalist who needs to acquire specific information about different topics during his/her professional activity. Or translators, who need both concise explanations of concepts and cross linguistic correspondences in order to understand specialized texts and translate them. Dictionaries can offer, in fact, proper assistance in a wide variety of different occasions, provided that they are reliable and efficient tools. The enormous inventory of specialized online dictionaries counts already reference works for top professionals in one field, like the authoritative *The New Palgrave Dictionary of Economics*, but also different hybrid[1] tools addressed to school children, like the entertaining *Math Spoken Here!*, which has been conceived to assist in learning and homework activities.

[1] For the concept of hybridization in electronic lexicography, see Granger 2011.

Surfing the Web it is possible to experience the tremendous amount of specialized dictionaries that are available for the most different fields. Compared to these resources, the number of general language vocabularies is but a few drops in the ocean. This state of affairs is however unsurprising, since similar disproportions were the rule in the paper dictionary era (Tarp, 2010), when vocabularies were not so easily accessible and one could not directly experience the real composition of the lexicographical production. The availability of these resources on the Internet has however overturned the proportion between the user, who is in need of lexicographical assistance, and the number of specialized resources he can consult, thus causing such an information overload that the user is either forced to resort to one of the usual *Wikipedia* pages, or to abandon the search completely. In both cases the user is stressed by the demanding activity of finding a source of information, rather than solving his/her information voids.

3 Solutions for integrated information access

Information overproduction on the Web has become one of the tasks of electronic lexicography since the advent of the first metalexicographical sites, called 'dictionary collections' (Engelberg and Müller-Spitzer, 2013), offering lists of links to different dictionaries. This practice has rapidly evolved into steadier solutions that have served also the opposite aim of a controlled integration of lexicographical data, made possible by the 'dictionary portals' (Engelberg and Müller-Spitzer, 2013) of well-established publishing houses, which have implemented the integration among their vocabularies in order to better meet the information needs of their users. In the *Pons* or *Cambridge* dictionary sites, for example, it is possible to access different vocabularies by filling in a single search mask and selecting the desired resource from a menu.

According to Engelberg and Müller-Spitzer (2013), dictionary portals "have followed [the] course from the single lexicographic product to the general lexicographic information service" that was predicted by Arnold (1979) and Kay (1983) as far as thirty years ago, thus creating a new type of dictionary. The possibility to cross-link well-structured informative resources, such as dictionaries, has in fact broaden the possibility of users to be informed promptly, by querying a single search engine that gives access to many dictionaries.

The right of ownership to the inventoried dictionaries is one of major restrictions determining the kind of access to the lexicographical information, thus influencing the portal typology. In the classification proposed by Engelberg and Müller-Spitzer (2013), dictionaries issued by the same publishing house may form 'integrated dictionary nets', if every vocabulary has been compiled with "a common concept of data modelling and structuring", thus allowing users to retrieve lemmata with similar properties from the different dictionaries inventoried, such as in the OWID. On the contrary, portals having no rights of ownership to the dictionaries, called 'dictionary collections', generally offer simple lists of links to external resources. Only a few of them are also provided with query systems that carry out searches in the lemma lists or in the whole text of the inventoried resources (see *OneLook*).

3.1 The *WLR* database assessment system

In addition to the types listed by Engelberg and Müller-Spitzer (2013), the *WLR* site increases the typologies of 'dictionary collections' by offering inventories of vocabularies that have been evaluated on the basis of the kind of data they contain (Caruso & De Meo, 2014). The assessment is carried out by a multi-parametric searchable database, which inventories dictionary features and assigns scores in order to display lists of resources that are more suited for two different types of parameters. It is in fact possible to search for dictionaries assisting with specific tasks, or 'lexicographical functions' that the dictionary should be able to fulfill (Tarp 2008), like acquiring new knowledge on a specific topic, solving communicative issues, or giving assistance with translations or learning tasks. These parameters can be set in *WLR* database by choosing the corresponding option in the 'Kind of assistance' box of the search form. Additionally, the user can set his/her level of expertise in the specialized field considered, and thus select the layman, semi-expert or expert profile in the 'Expertise level' box.

The rating system used in the *WLR* site is intended to increase the effectiveness and efficacy of portals, making dictionary collections less time wasting and more useful also for the less experienced dictionary users, since they avoid the display of "long lists" that show "results from trustworthy sources and downright amateurish concoctions all mixed up" (de Schryver 2003: 157). The evalua-

tion system relies in fact on the presence or absence of 58 types of data, addressing all the component parts of dictionaries (Caruso & De Meo, 2014): from the host site and the general organization (or macrostructure), to the mediostructure and microstructure, for which both linguistic and encyclopedic data are taken into consideration. Additionally, explicit guidelines are followed for the score assignment system: characterizing data for a specific parameter receive one or two points, according to their degree of relevance. Negative scores (-1, -2) are instead given to contradictory data. Similarly, each lexicographical parameter considered ('Kind of assistance' and 'Expertise level') can reach the same maximum score: for example, the different types of users may have no more than 24 points. In the meanwhile, for contradictory profiles, such as laymen and experts, the score distribution cannot be the same.

All this things considered, one can affirm that the *WLR* site aims to support different types of users decreasing the information overload that occurs while consulting rich inventories of non-integrated resources, such as dictionary collections sites. Additionally, the *WLR* rating system is in line with the parameters identified by Swanepoel (2008; 2013) to carry out dictionary evaluations that are scientifically grounded, i.e. assessments that explicitly state the analytic principles they use and the way these are applied, together with instructions to measure the compliance or non-compliance to these same principles.

Additionally, the portal wires together fragments of the huge repository of specialized knowledge available on the Internet (Caruso 2014), hosting dictionaries of around 60 different fields, such as oenology, mathematics and medicine.

4 How to make effective searches

Recent studies have underlined that electronic dictionaries are special types of information systems (Tarp, 2008; Bothma, 2011; Gows, 2011; Heid, 2011) and evaluative parameters borrowed from the Information Science are used in the literature on electronic lexicography topics. In particular, the quality of one dictionary can be assessed on the basis of its usefulness for a task completion, like finding a specific collocate while writing a text. Therefore, the dictionary is considered to be effective if it provides "the right data and the right amount of data to the user" (Heid 2011: 290). On the contrary, it is efficient if gives quick access to the data needed.

The *WLR* database developed so far assures that the search for a dictionary is less time wasting for the user but it does not guarantee that the data provided by one dictionary are correct or correctly stated. Contrarily, the quality of data is always paramount, and users' searches would be more effective if they could avoid to consult vocabularies whose data are unreliable.

For example, the following Spanish oenological dictionary (*Infoagro.com - Diccionario del vino*) explains that 'ácido' is a "green wine" whose colour seems to be a consequence of a bed fermentation:

> [1] "Ácido: Vino verde. Producto de una mala fermentación maloláctica, una uva en mal estado o recolectada antes de tiempo."

On the contrary, many other dictionaries explain the same term as denoting a sour wine, or a wine that is high in acidity, like in following entry (*Diccionario del vino.com*):

> [2] "Ácido: 1.- Vino cuya acidez sobrepasa la media de la región. La acidez puede ser debida a un exceso de ácidos organicos o a un desequilibrio entre los sabores del vino.
> 2.- Vinos con PH inferior a 3,2"

In order to carry out more efficient searches using the current release of the *WLR* database, one can look for dictionaries compiled exclusively by authoritative institutions, thus restricting the search to 'Institutional' and 'Specialized' host sites, two features that users can select in the database search form. However, even the dictionaries edited by the most authoritative institutions offer examples of bad explanations that can be misleading for the user, or even difficult to interpret (Caruso & De Meo, 2014). For example, the *Talking Glossary of Genetics*, published by the National Human Genome Research Institute, in the *Chromosome* definition explains that: "Humans have 23 pairs of Chromosomes (…), and one pair of sex chromosomes, X and Y". Stated this way the definition is incorrect, since only male humans have an XY pair of chromosomes, while females have an XX pair. Effective lexicographical definitions should obviously provide more complete descriptions and should avoid incorrect generalizations like this.

Assessing data quality poses however many methodological and theoretical problems regarding the

terms and the definition features that must be rated (see Caruso & De Meo, 2014) by the system. For example, the number of the assessed lemma must remain the same despite the number of dictionary entries? Which definition features are suited to estimate whatever concept belonging the specialized fields as different as, for example, figurative arts and finance? Furthermore, at least one expert for each specialized field considered should verify the information provided, which is probably the most serious obstacle to future developments of the project. However, a completely different solution has been imagined, as will be shown in a moment.

4.1 The database as a data validation tool

The *WLR* database has been conceived as a flexible tool that allows its administrators to add or change labels in the three component parts that make up the repository system, which are called 'categories', 'features' and 'rating system'. The first component, or 'category', lists the types of inventoried linguistic resources: only dictionaries have been assessed so far, but other supportive instruments to solve linguistic issues could be added to the database, like corpora or grammars. To each category the administrator assigns different descriptive features, which is the second component of the rating system, and can be both binary or multivalve. The 'dictionary' category has 58 feature (see Caruso, 2014 for a complete list), some of them can only be present or not, thus are binary, like *Cultural Notes*, others are multivalue and thus need further specifications, like the *Kind of Dictionary*, which must be set choosing among different choices: *Monolingual dictionary*, *Monolingual word list*, *Multilingual dictionary*, *Multilingual word list*, *Plurilingual dictionary*[2]. Lastly, grades are assigned to each of these values according to the methodology described above. The administrator can decide to set different evaluative parameters for each category taken into account: for example, if grammars were added to the repository, the language proficiency level could be a suitable evaluation parameters for it.

Once however that the grades distribution has been decided, the database assigns points automatically and independently from any actions performed by the compiler of the evaluation forms, who can set only the values of the different features. Likewise, if the score assignment is changed, the inventoried dictionaries will immediately change their evaluations. The automatization of grades assignment guarantees no errors in the final score computation, however, the selection of values that describe the dictionary features are of crucial relevance for the accuracy of the evaluation.

Under this respect, the inventoried resources must be analysed carefully, because most of the times specialized online dictionaries lack strict lexicographical organization and display different data types unsystematically: for example, basic information on the word form might be given exclusively in some of the entries of one dictionary, independently of any significant paradigmatic variation of the language considered. For similar cases, the compiler must set the 'sometimes' value in the corresponding feature of the evaluation form, and the record of the data that are sporadically given by the dictionary will make the evaluation procedure more reliable.

Actually, the current development of the project is improving the existing database components with an additional part that keeps track of where unsystematic data, like those mentioned above, are present in the dictionary. This addition will make the assessment procedure extremely reliable, since the less evident features can be registered, making the evaluation accuracy easily verifiable.

With this new database component, the evaluation forms will be fillable by anyone and the *WLR* database will become a collaborative portal. This, hopefully, will make the number of the inventoried resources increase, and it will offer other additional developments.

While compiling the forms, in fact, users could also contribute to verify the quality of the data provided, signalling for each dictionary feature if any wrong information is given. For each inconsistency the user should indicate one alternative data and the source of information from which this was driven. On the other hand, the database will offer warning signals that indicate the presence of problematic data within one dictionary.

Acknowledgements

We wish to thank Gianluca Monti for managing the first version of the *WLR* database and site.

The present research has been sustained by academic grants from the University of Naples 'L'Orientale'.

[2] For the concept of Plurilingual Dictionary, see Caruso, 2011.

References

Arnold, D. I., 1979, "Synonyms and the College-Level Dictionary", Dictionaries, 1: 103–12.

Bothma, T.J.D, 2011, "Filtering and Adapting Data and Information in an Online Environment in Response to User Needs", in Fuertes-Olivera, P.A., Bergenholtz, H. (Eds.), 71-102.

Carr, M., 1997, "Internet Dictionaries and Lexicography", International Journal of lexicography, 10/3: 209-230.

Caruso V., 2011, "Online specialised dictionaries: a critical survey", in Kosem I., Kosem, K. (eds.) *Electronic lexicography in the 21st century: new applications for new users. Proceedings of eLex 2011*, Ljubljana: Trojina, Institute for Applied Slovene Studies, 66-75.

Caruso V., 2014, "A Guide (not only) for Economics Dictionaries", Hermes – Journal of Language and Communication in Business, 52: 75-91.

Caruso, V., De Meo, A., 2014, "A Dictionary Guide for Web Users", in Abel, A., Vettori, C. & Ralli, N. (eds.), *Proceedings of the XVI EURALEX International Congress: The User in Focus*, Bolzano: EURAC, 1087-1098.

De Schryver, G. M., 2003, "Lexicographers' Dreams in the Electronic-Dictionary Age", International Journal of Lexicography, 16/ 2: 143-199.

Engelberg, S., Müller-Spitzer, C, 2013, "Dictionary Portals", in Gouws, R. H., Heid, U., Schweickard, W., e Wiegand, H. E. (eds.), *Dictionaries. An international encyclopedia of lexicography. Supplementary volume: Recent developments with special focus on computational lexicography*. Berlin/New York: de Gruyter, 1023-1035.

Fuertes-Olivera, P.A., Bergenholtz, H. (Eds.), 2011, *e-Lexicography: The Internet, Digital Initiatives and Lexicography*, London, New York: Continuum.

Gouws, R.H, 2011, "Learning, Unlearning and Innovation in the Planning of Electronic Dictionaries", in Fuertes-Olivera, P.A., Bergenholtz, H. (Eds.), 17-29.

Granger, S., 2012, "Introduction: Electronic Lexicography from Challenge to Opportunity", in Granger, S. & Paquot, M. (Eds.), Oxford: OUP, 1-11.

Heid, U. 2011. 'Electronic Dictionaries as Tools: Towards an Assessment of Usability.' In P. A. Fuertes-Olivera and H. Bergenholtz (eds.), 287–304.

Kay, M., 1983, "The dictionary of the future and the future of the dictionary", in Zampolli, A. & Cappelli, A. (eds.), *The Possibilities and Limits of the Computer in Producing and Publishing Dictionaries: Proceedings of the European Science Foundation Workshop*, Pisa: Giardini Editori, 161–74.

Swanepoel, Piet, 2008, "Towards a Framework for the Description and Evaluation of Dictionary Evaluation Criteria", Lexikos, 18: 207-231.

- 2013, "Evaluation of dictionaries", in Gouws, R. H., Heid, U., Schweickard, W., e Wiegand, H. E. (a cura di), *Dictionaries. An international encyclopedia of lexicography. Supplementary volume: Recent developments with special focus on computational lexicography*. Berlin/New York: de Gruyter, 587-596.

Tarp, S., 2008, *Lexicography in the borderland between knowledge and non-knowledge*, Tübingen: Niemeyer.

- 2010, "Beyond Lexicography: New Visions and Challenges in the Information Age", in Bergenholtz, H., Nielsen, S. & S. Tarp (eds.), *Lexicography at a Crossroads. Dictionaries and Encyclopedias Today, Lexicographical Tools Tomorrow*, Bertlin et.: Peter Lang, 17-32.

Online Dictionaries and resources

Cambridge Dictionaries, http://dictionary.cambridge.org/it/, accessed July 2016.

Diccionario del vino.com, http://www.diccionariodelvino.com/index.php/letra/a/.

Infoagro.com - Diccionario del vino, http://www.infoagro.com/viticultura/diccionario/diccionario.htm, accessed July 2016.

Math Spoken Here! An Arithmetic and Algebra Dictionary, http://www.mathnstuff.com/math/spoken/here/, accessed July 2016.

OneLook, www.onelook.com, accessed July 2016.

OWID (*Online-Wortschatz-Informationssystem Deutsch*), http://www.owid.de/, accessed July 2016.

Pons, www.pons.eu, accessed July 2016.

Talking Glossary of Genetics, https://www.genome.gov/glossary/, accessed July 2016.

The new Palgrave dictionary of economics, Basingstoke and New York: Palgrave Macmillan, http://www.dictionaryofeconomics.com, accessed July 2016.

Web Linguistic Resources (*WLR*), www.weblinguisticresources.org, accessed July 2016.