Turning the corpus into a functional component of the dictionary: The case of the Oenolex Wine Dictionary

Patrick Leroyer*

Aarhus University, Aarhus School of Business and Social Sciences, Department of Business Communication, Centre for lexicography, Jens Kr. Skous Vej 4, 8000 Aarhus C, Denmark

Abstract

Lexicographical corpora for dictionary making are normally used as an empirical base of the dictionary. They provide highly valuable lexical material, patterns, attestations of use, etc. However, once exploited by the lexicographer, they remain hidden to the user. The question thus is: Can a corpus become even more useful not only to the lexicographer, but also to the user? How can a corpus become a functional component of the dictionary? I will address these questions by outlining the methodological issues related to the design and exploitation of an oral corpus for the construction of Oenolex, a dictionary of wine tasting. In Oenolex, the corpus is not simply the empirical base of the dictionary: it is lexicographically structured so that it can be accessed by the user. Oenolex also shows how oral corpora can be an asset for specialised dictionary projects, particularly in domains in which terminology is subject to a high degree of variation. This in turn calls for an even closer cooperation with experts; I will therefore argue, on the basis of the Klosa model of lexicographical phases (2013), that experts from the field of specialised communication and knowledge covered by the dictionary are truly needed in almost all phases of the dictionary process.

© 2015 The Authors. Published by Elsevier Ltd.

Keywords: lexicographically structured corpora; oral corpora; wine dictionary; wine tasting terminology; cooperative lexicography

1. Introduction

Lexicographical corpora for dictionary making are normally used as the empirical base of the dictionary. They provide highly valuable lexical material, patterns, attestations of use, etc., but once exploited by the lexicographer,
they normally remain hidden to the user. The question thus is: can a corpus gain a more important function andecome even more useful not only to the lexicographer, but also to the user? In other words, how can a corpus be
turned into a functional component of the dictionary? I will address these questions by outlining the methodological
issues related to the design and exploitation of an oral corpus for the construction of Oenolex, a cooperative
dictionary of wine tasting. I will also argue, on the basis of the Klosa model of lexicographical phases (2013), that
experts from the field of specialised communication and knowledge covered by the dictionary are truly needed in
almost all phases of the dictionary process. The Oenolex dictionary project (Leroyer, 2011, 2013a, 2013b, 2014;
Leroyer and Høy, 2014; Gautier and Leroyer, 2014) aims at the construction of a professional, online dictionary of
wine tasting for the French wine industry of Burgundy. It is organised and managed as an international lexicographic
co-operation between the University of Burgundy in Dijon (France) and Aarhus University (Denmark). The
dictionary is commissioned by the BIVB, the branch organisation of the Burgundy wine industry in France, and the
French Region of Burgundy. What the BIVB really has in mind is an information tool supporting the promotion of
the dissemination of knowledge of their wines. Therefore, the lexicographic team decided at an early stage, in
agreement with the BIVB, to develop a lexicographic information tool aimed at the communicatively and
cognitively oriented information needs of two categories of users engaged in teaching and learning wine tasting in
various situations, including wine promotion:

1. Expert users, namely the wine experts (oenologists) teaching wine tasting courses at the Burgundy wine
   school and the wine producers or wine retailers planning and delivering promotional communication of
   Burgundy wines

2. Non-expert users, namely the students attending wine tasting courses at the Burgundy wine school

The Oenolex project was launched back in autumn 2013, and so far the conceptual phase and the data acquisition
phase have been completed, while data processing and database structuring are still in progress. The experts of the
BIVB have been closely involved in all phases, as they have decided on the specifications of the dictionary concept,
and participated actively in the generation and acquisition of the lexicographic data. The data come from a corpus of
BIVB promotional documents and from an oral corpus of recordings of authentic wine tasting interactions between
teachers (wine experts) at the Burgundy wine school, which is run by the BIVB, and students (non-experts) taking
these courses.

2. Acquiring and structuring the data

2.1. Corpus and sub-corpora

Oenolex makes use of both written and oral data sources. On the one hand, Oenolex exploits a text corpus of
promotional and pedagogical online wine records and wine tasting guides. On the other hand, Oenolex exploits, in
close cooperation with the BIVB and other actors of the profession (wine makers, fair organisers), an oral corpus of
wine tasting interactions specially recorded for the purpose. One reason for using oral corpora is due to the fact that
the major part of professional interactions in wine tasting takes place in oral production situations. A nother reason
lies in the dynamic nature and fuzziness of wine tasting terminology and wine tasting descriptors. Recent research
(Gautier and Hohota, 2015) shows that the terminology of wine tasting is subject to a high degree of variation, as
terms and expressions are co-constructed in and by the discourse of wine during wine tasting sessions. For the
making of the prototype, we have produced 12 hours of recordings of wine tasting situations, including teacher
presentations, wine tasting workshops, wine tasting fairs, and wine tasting at wineries. It is our plan in the future to
include podcasts of wine tastings presentations, and also data generated by online surveys among samples of
consumers of Burgundy wines and among wine experts. The oral corpus includes two sub-corpora, a didactic corpus
(recorded at the wine school) and a promotional corpus (recorded at wine estates and wine fairs). Although quite
similar on the whole, these corpora show some interesting differences in degrees of interactions; table 1 below
presents in a simplified form the main characteristics of the two sub-corpora:
Table 1 – The two oral sub-corpora of Oenolex (adapted from Gautier and Hohota, 2015)

| Criteria                              | Wine school          | Wine estates and wine fairs          |
|---------------------------------------|----------------------|-------------------------------------|
| Type of interaction                   | public               | semi-public/private                  |
| Degree of familiarity between actors  | very low             | variable, evolves over time          |
| Part of emotion towards object of discourse | high               | high                                |
| Integration of situation and action   | high                 | high                                |
| Cooperation between actors            | high                 | high                                |
| Degree of proximity                   | sequential/present   | present                             |
| Degree of dialogicity                 | medium               | low                                 |
| Degree of spontaneity                 | medium               | low                                 |
| Interrelation between situation and object of discourse | high               | high                                |

2.2. Database structure

As Oenolex is conceived as a professional dictionary of wine tasting of Burgundy wines exclusively, it is designed so that its user and commissioner, the BIVB, can systematically record, store, and retrieve specific data associated to the tasting of their wines in various tasting sessions. The database includes the following data sets:

- a pragmatic register (table) of all wine tasting sessions with records of the specific wines tasted, place, date, participants, and evaluations assigned to the wines by experts and non-experts
- a multimodal (sound, text, images), encyclopaedic register (table) of all tasted AOC-wines with records of the detailed specifications of every unique AOC (region, category of AOC, name, etc.)
- a multimodal (sound, text, and image files) terminological register (table) of all wine tasting terms and expressions used during wine tasting sessions containing lemma records, grammatical and phraseological information, examples linked to the corpus, sources, and external links
- an underlying digitalized and lexicographically structured corpus of recordings of wine tasting sessions and sequences of wine tasting sessions. Wine tasting sessions are split up into sequences determined by identification of the unique AOC being tasted, and by the sequential phases of wine tasting (macro-sequencing), and by various discourse markers and knowledge contexts (micro-sequencing) for the extraction of good lexicographic examples (see also below, section 2.5, criteria for extraction of good examples)

The three registers and the corpus are linked together through a network of relations between data sets and data fields (the limited amount of place allowed in this contribution doesn’t permit to show these), which makes it possible to make targeted searches on:

- wine tasting terms and expressions as they are used in situated contexts by different actors
- circumstances under which these terms and expressions have been actually used in interaction: which terms and expressions are assigned to which wines, by whom, in what circumstances, where and when?

The design of the data base as a mash up of pragmatic, encyclopaedic, and lexical data is primarily oriented towards the satisfaction of cognitive functions in which the users – the experts of the BIVB – use the dictionary because they need to acquire knowledge on the discourse of wine tasting in authentic situations, and how this discourse acts upon the terminology of wine tasting, how it is created, communicated, understood (or misunderstood), how it develops, etc. The discourse of wine tasting also features interesting didactic elements (wine
tasting teaching at the wine school in order to learn about wine) and promotional elements (wine tasting at the winery in order to sell wine), which can be used by the BIVB to develop new courses or promotional campaigns.

2.3. AOC-data and wine tasting data (not illustrated here)

For each single wine, the AOC data include the following records: full name of AOC, category of AOC, grape variety, climate and locality, village, year, producer, region, wine making process, visual evaluation, olfactive evaluation, gustative evaluation, and general evaluation (not illustrated here for space reasons).

Every single wine tasting session includes the following data records: tasted AOCs, date of tasting, place of tasting, wine tasters present, as well as all 4 types of evaluations (visual, olfactive, gustative, and general) from both experts and non-experts involved (not illustrated here for space reasons).

2.4. Lemma selection and crafting of definitions

Lemmas are manually selected from the corpus of wine tasting evaluations. The corpus is a collection of recordings of wine tasting sessions that have been transcribed using speech-to-text software, followed by manual cleaning. The lemmas represent all the tasting characteristics attributed to the wines by both experts and non-experts. Other terminological units which are not related to wine tasting proper, e.g. chemical terms (glycerol) or viticultural terms, are excluded from lemma selection. Lemma records include the following data fields: lemma, grammatical information on lemma, definition, source of definition, collocations, examples, source of examples as well as 3 specific BIVB data fields: oenological note, encyclopaedic note, illustration.

Definitions, generally quite short, are terminologically crafted according to genus proximum and differentia specifica and make use of the conceptual features chosen by the BIVB in their reference documents, such as wine tasting guides and protocols (Le guide des arômes des vins de Bourgogne being used as a reference for the definition of wine tasting descriptors and L’art de la dégustation en Bourgogne for other wine tasting terms). Definitions are submitted to experts of the BIVB for validation. An example of a definition is given below for the lemma animal:

animal famille d'arômes faisant référence aux senteurs qui sont liées aux animaux telles que la fourrure, le cuir, le musc ou l'ambre, et qui peuvent être identifiés lors du nez et/ou de la rétro-olfaction

[= animal fragrance family referring to smells connected to animals such as smells of fur, leather, musk or amber, that can be identified in olfaction and retro-olfaction phases]

2.5. Selection of good lexicographic examples

As we look for didactic quality, the selection of examples includes sequences in which the purpose of the teacher/student interaction is to redefine a term, correct mistakes or misunderstandings. Examples are also taken from operative instructions sequences, in which the teacher explains in a dialogical form how to proceed with the wine tasting. In both corpora, sequences are largely pre-established, as they follow the canonic, chronological, and sequential phases of wine tasting: presentation of the wine → visual aspects → nose → taste → general impressions and concluding remarks. This feature helps us delimitate the corpus and assign the data to the relevant data fields in the database (visual, olfactive, gustative, general impression, as shown in table 3). Stereotyped contributions (such as “au premier nez on est sur...” = [at first nose we are on...], “au deuxième nez on évolue...” = [at second nose we move to...]) help us select collocations, which include a great number of support verb constructions. Questions and riddles (“Le vert indique la jeunesse ou l'origine?” = [Does green indicate youth or origin?]) are also effective indicators of important pedagogical moves and are valuable candidates to the selection of good examples.
3. User interface: Search modes and display of articles

The OENOLEX user interface provides two basic search modes to its users, both being activated by function buttons: the AOC search mode (= AOC button on user interface) and the wine tasting term search mode (= TERM button on interface). In the underlying data base structure, every single AOC is related to the terms that have been used to describe it, as well as to the specific wine tasting sessions under which the terms have been used.

It is also possible for the user to refine his/her search (in order to show less) and adapt the display of results by filtering out the data according to the following search options: region, wine process, AOC category (grand cru, premier cru, village, regional) as far as the AOC-search is concerned, and place of wine tasting session (wine schools, wine auction, wine fair, etc.), date of wine tasting session, and types of wine tasters present under wine tasting (wine teacher, oenologist, producer, etc.) as far as TERM-search is concerned. The presentation of results includes 5 different displays of articles, 2 for display of terms, and 3 for display of AOCs.

3.1 Display of AOC-articles

AOC-articles include all data related to the unique AOC, as well as a list of all wine tasting sessions in which this AOC has been tasted, see table 2 below:

Table 2: AOC ID-article

| AOC          | Chitry, 2010, Domaine Olivier MORIN |
|--------------|-------------------------------------|
| Catégorie    | village                             |
| Cépage       | Chardonnay                          |
| Climat & lieu-dit | Chitry                        |
| Commune      | Chitry                              |
| Millésime    | 2010                                |
| Producteur   | Domaine Olivier MORIN               |
| Région viticole | Chablis                        |
| Vinification | blanc                               |

Wine tasting-list: 07 novembre 2013, Ecole des vins de Bourgogne

Clicking on the wine tasting session(s) from the list (in the above example, the list only contains one item) in which this specific AOC has been tasted will display the evaluation article of the AOC, see table 3 below, in which all lemmatised terms are hyperlinked. Clicking on one of these opens the corresponding article. Links are also provided to the corpus, to sound and video files, and to selected, external web page resources in line with Fuertes Olivera (2012) who recommends the use of the Internet as a lexicographical (re)source.

Table 3: AOC evaluation article

| AOC          | Chitry, 2010, Domaine Olivier MORIN |
|--------------|-------------------------------------|
| Date         | 07 novembre 2013                     |
| Lieu         | Ecole des Vins de Bourgogne, 6 Rue du 16ème Chasseurs, 21200 Beune |
| Évaluation visuelle expert | L à on a une robe avec de jolis reflets verts. |
| Évaluation olfactive expert | On sent le citron, le pamplemousse, on peut dire qu’il est un peu exotique. Le deuxième nez, il paraît plus fort, plus cristaux, mais moins citronné, ça change un peu. |
| Évaluation gustative expert | En bouche, vous allez sentir acidité et sucrosité. En bouche, il y a aussi une certaine minéralité, ça sent un peu le côté caillou. |
| Évaluation générale expert | C’est un vin qui peut se conserver encore un peu. Sur la fraîcheur, sur la jeunesse, bon à déguster maintenant, |
3.2. Display of term articles and links to AOC articles

Term articles, see table 4 below, contain all the fields already mentioned above: term definition, grammatical information, list of collocation, list of examples, and type of source from which the term has been extracted (corpus). Below the term article are listed all the AOCs in which the term, here pamplemousse, has been used. Clicking on an item of the AOC list opens the corresponding AOC-article at the position of the lemma in its context.

Table 4: Term article

| Term     | pamplemousse |
|-----------|--------------|
| Grammaire | nom masculin |
| Définition| arôme catégorisé dans la famille d'arômes "fruits frais, agrumes", faisant référence aux senteurs du pamplemousse, identifié lors du nez et/ou de la rétro-olfaction |
| Collocation| être sur le pamplemousse |
| Exemple | Donc, on était sur des fruits à chair blanche et le côté un petit peu citronné, mandarine, un petit peu pamplemousse, ce genre de choses. |
| Type de source | transcription du 7 novembre 2013, École des vins de Bourgogne, Beaune |
| AOC list: | Bourgogne Chitry, Chitry, 2010, Domaine Olivier MORIN |
|           | Mercurey blanc, Chardonnay, St. Martin-sous-Montaigu, 2012 |

4. Cooperative construction of OENOLEX: The participation of end-users and various experts

As stated by Fuertes Olivera (2013), applying new technology in e-lexicography constitutes a continuing challenge which also raises the question of participation of various experts and of end-users in the construction of the dictionary. Klosa (2013) has designed a work process model for the construction of e-dictionaries describing the various work phases and their content, from preparation of the dictionary to afterlife upon publication. Based on this model, we propose a modified model, see figure 1 below. Some of the descriptions of the various phases have been left out for the purpose of this article, as the objective of this modified model is to show to what extent dictionary users and experts of the field covered by the dictionary should be connected or not unto the different phases of the construction process. Based on our experience until now in the construction of Oenolex and in the preparation of the prototype, we are in a position to formulate some general tenets and recommendations. The model shows that the active involvement of users and experts in all phases of the work flows in the construction of the dictionary. The model also specifies the conditions of their involvement: to what degree should experts or users be involved or not.
In phase 1, preparation of the concept, the experts should always be involved in decision making regarding the specific design of the lexicographic concept, particularly the degree of specialisation, the identification, and definition of the target users and their profile, and the functions to be implemented, i.e. the foreseen situations in which the dictionary is supposed to help its intended users: documentation and knowledge acquisition needs, needs for help in connection with specialised translation, text editing or technical writing, etc. In the case of Oenolex, the BIVB has been the prescriber and has subsequently taken part in all decision making.

In phase 2, acquisition, both end-users and experts can be involved depending on the genuine purpose of the dictionary. In the case of Oenolex, both groups are needed because the data is generated in cooperation with experts, and include a written expert corpus of specialised documents as well as an oral, mixed corpus of recordings of specialised interactions between experts and non-experts, which subsequently are subject to a thorough discourse analysis. Thus, in Oenolex, data acquisition is the direct continuation of the preparation phase.

Phase 3, computerisation, normally does not require the participation of either end-users of the dictionary or experts of the field, as the lexicographic team will normally manage this part of the work flow according to the database and corpus technology used by the lexicographic team, cf. for instance Nielsen and Akmind 2011. In the case of Oenolex, an IT-expert is needed, and computerisation is supported by programmers for all technical aspects concerning the database, including the dictionary writing system.

In phase 4, processing, experts will rarely be needed. In the case of Oenolex, experts will occasionally be consulted in the tagging process of audio sequences in order to validate the lexicographic quality of some examples of interactions between experts and non-experts – or reject these if necessary.

In phase 5, analysis, experts are always needed. This is actually a sine qua non condition of specialised lexicography. The lexicographer is normally an expert in lexicography, but rarely an expert of the subject-field covered by the dictionary. Only experts can craft definitions of concepts and relations between concepts that have
the necessary lexicographic authority (cf. Hashimzade, Myles, and Myles, 2014). In Oenolex, definitions and explanations are submitted to the experts of the BIVB.

In phase 6, testing prior to release, the involvement of both users and experts is highly advisable. Only by then can their involvement in previous phases be assessed by the lexicographic team in the ongoing construction process of the dictionary; only by then can the lexicographic team make the necessary changes or adaptations in the work processes. It is expected that we will have to make changes and adaptations after the release of the prototype.

In phase 7, afterlife, both users and experts are always needed because the construction of the online dictionary is a recurrent, continuous loop process, much in the same way as in phase 6 above. User-surveys can also provide valuable feedback on the user-friendliness of Oenolex, so users will definitely be needed in the future developments and updating of the dictionary.

5. Conclusion: The corpus at the very heart of the dictionary and cooperative lexicography

Our experience with Oenolex so far clearly confirms that oral corpora can become a valuable asset in specialised, professional dictionary projects dealing with dynamic fields of knowledge, because oral interactions, although not as dialogic as expected, play an important role in the (de-) and (re)construction of knowledge of the profession. Our experience also clearly confirms that the corpus should not simply be treated as an empirical substratum for the dictionary, but can be turned into a full-fledged functional component of the dictionary and put at the very heart of the dictionary so to speak. This necessitates a specific lexicographic structuring of the database linking lexicographic articles and examples from the corpus closely together, and consequently the programming of the user search interface accordingly. Corpus and dictionary are not to be seen as separate entities, but as the two complementary sides of a modern linked data information system which can be logically and effectively integrated.

Finally, working with Oenolex has told us that the lexicographic management team should cooperate with a panel of experts almost all the way, and not only, as commonly accepted today by specialised lexicographers, to check on and validate definitions of specialised items from the lemma list that have been selected and crafted by the lexicographer. In todays’ world, professional words in professional discourses can no longer be separated from things, actions and events in the world, and professional, expert knowledge is needed in almost all phases of the lexicographic work. In other words, specialised lexicography is a truly cooperative and interdisciplinary discipline taking advantage of mixed methods of its own to adapt, in every single case, the treatment and the presentation of the lexicographic material to the genuine information needs of its intended users in the foreseen use situations.

References

Fuertes Olivera, P. (2012). Lexicography and the Internet as a (Re-)source. Lexicographica 28, 49 - 70.
Fuertes Olivera, P. (2013). E-lexicography: The Continuing Challenge of Applying New Technology to Dictionary Making. In H. Jackson (Ed.), Bloomsbury Companion to Lexicography (pp. 323-340). London/New Delhi/New York/Sydney: Bloomsbury Academic.
Gautier, L., and Hohota, V. (2015). Construire et exploiter un corpus oral de situations de dégustation: l’exemple d’Oenolex Bourgogne. Studia Universitatis Babes-Bolyai-Studia Philologia (pp. 157-173).
Gautier, L., and Leroyer, P. (2014). Construction, communication, représentation et réappropriation des discours vitivinicoles dans un ‘nuancier’ lexicographique en ligne. In C. Condei et al. (Eds.), Situations professionnelles, discours et interactions en traduction spécialisée. Berlin: Frank und Timme, GmbH Verlag für wissenschaftliche Literatur [in press].
Hashimzade, N., Myles, G. A., and Myles, G. D. (2014). Can Authority be Sustained while Balancing Accessibility and Formality? HERMES 52, 11 - 24.
Klosa, A. (2013). The Lexicographical Process (with Special Focus on Online Dictionaries). In R.H. Gouws, U. Heid, U., W. Schweickard, and H. E. Wiegand (Eds.), Dictionaries. An International Encyclopaedia of Lexicography. Supplement Volume: Recent Developments with Focus on Electronic and Computational Lexicography (pp. 517-524). Berlin, Boston: de Gruyter.
Leroyer, P. (2011). Change of Paradigm in Lexicography: From Linguistics to Information Science and from Dictionaries to Lexicographic Information Tools. In P. A. Fuertes-Olivera and, H. Bergenholtz (Eds.), E-Lexicography: The Internet, Digital Initiatives and Lexicography (pp. 121-140). London, New Delhi, New York: Routledge, Internationaler Verlag der Wissenschaften.
Leroyer, P. (2013a). New Proposals for the Design of Integrated Online Wine Industry Dictionaries. Lexikos, 23, 209 - 227.
Leroyer, P. (2013b). Putting Words on Wine: OENOLEX Burgundy, New Directions in Wine Lexicography. In D. A. Kwary, N. Wulan, and L.
Musyahda (Eds.), Lexicography and Dictionaries in the Information Age: Selected Papers from the 8th ASIALEX International Conference (pp. 228-235). Airlangga University Press, Airlangga.

Leroyer, P. (2014). La lexicographie du vin: état des lieux théorique et monofonctionnalité modulaire. In I. Rousseau-Jacob, and L. Gautier (Eds.), Figures et images dans le discours sur le vin en Europe. Peter Lang [in press].

Leroyer, P., and Gautier, L. (2014). OENOLEX Bourgogne. Construction, communication, représentation et réappropriation des discours vitivinicoles dans un nuancier lexicographique en ligne. In Situations professionnelles, discours, interactions : vers une didactique de la traduction. Frank & Timme GmbH Verlag für wissenschaftliche Literatur [forthcoming].

Leroyer, P., and Høy, A. (2014). OENOLEX Bourgogne: Lær at sætte ord på vin. Nyt vinordbogsconcept, nye veje for brancheordbøger. In R. V. Fjeld, and M. Hovdenak, Nordiska Studier Ilexikografi (pp. 287-303). Oslo: Nordisk Forening for Leksikografi.

Nielsen, S., and Almind, R. (2011). From Data to Dictionary. In P., and H. Bergen Holtz (Eds.), E-lexicography, The Internet, Digital Initiatives and Lexicography (pp. 141-167). London, New Delhi, New York, Sidney: Continuum.