Extracting (good) discourse examples from an oral specialised corpus of wine tasting interactions

Patrick Leroyer, Laurent Gautier, Hédi Maazaoui

Extracting (good) discourse examples from an oral specialised corpus of wine tasting interactions

Patrick Leroyer[1], Laurent Gautier[2], Hedi Maazaoui[3]

[1] Aarhus University, Jens Chr. Skous Vej 4, DK-8000 Aarhus
[2]Université de Bourgogne, 6 esplanade Erasme, BP 26 611, F-21066 Dijon Cedex
[3] Université de Bourgogne, 6 esplanade Erasme, BP 26 611, F-21066 Dijon Cedex

E-mail: pl@asb.dk, laurent.gautier@u-bourgogne.fr, Hedi.Maazaoui@u-bourgogne.fr

Keywords: oral corpus, Sonal® software, thematic tagging, extraction of discursive examples, discursive markers, interaction frames, segmentation

1. Compiling a professional dictionary of wine tasting interactions

This article outlines the semi-automated extraction of dictionary examples used in the compilation of a professional online dictionary of wine tasting. Named OenoLex Bourgogne, this dictionary was started to respond to the demand for a lexicographic information tool from the French wine industry of Burgundy, the *Bureau Interprofessionnel des Vins de Bourgogne* (BIVB)[1]. It has been developed by the University of Burgundy in Dijon (France) and the University of Aarhus, Centre for lexicography (Denmark). OenoLex is not a conventional language dictionary (Coutier, 2011) or terminological dictionary of wine (Bottlenotes, 2015) as its purpose is to systematically represent the discourse of wine tasting when experts and non-experts meet and interact in various kinds of wine-tasting situations in connection with the evaluation of specific wines. OenoLex is designed for expert users and provides the BIVB with valuable knowledge of what is actually being said about their wines when producers, teachers and consumers have an opportunity to meet[2]. In the data base we index all specific wines being tasted and connect them to the wine-tasting descriptors and description sequences at work in their actual evaluation.

2. What are good example sentences in OenoLex? Modified status

The purpose of good lexicographic example sentences is to illustrate the correct meaning (for text reception) and use (for text production) of the lemma in context, and

[1] The dictionary is co-financed by the BIVB (Bureau Interprofessionnel des Vins de Bourgogne) and by grants from the Regional Council of Burgundy.

[2] For a more detailed presentation of the dictionary see Leroyer (2013), Gautier & Hohota (2014), Gautier & Leroyer (2015), and Leroyer *and al*. (2015).

thus reinforce the information provided by the dictionary articles (meaning explanations, restrictions of use, collocations, etc.). The formal requirements for the quality of good lexicographic examples include among other things such primary factors as length (not too short, not too long), not too specific or too vague, free of corpus noise, typos, slang, and sensitive content (Michelfeit, 2015). In OenoLex, examples do not have to comply with these requirements as they share the fuzziness and natural variation of authentic interactions in (oral) professional discourse (Gunnarsson, 2009); they are assigned the status of main information unit in the dictionary and are not subordinated to the lemma in the same way as in conventional dictionaries, the function of lemmas being here to provide access to the structured examples of wine tasting interactions.

3. The corpus and its sub-corpora

At present[3], the corpus includes 12 recordings *in vivo* from speech interactions, the total length being 8hrs and 15min. Depending on context, the length of the individual recordings varies from less than 10min for the shortest one to 3hrs for the longest one. Despite length variation, there is no significant skewing in the data as all discourse types are extremely stereotyped. The recordings all deal with a number of specific appellations from Burgundy (Côte de Beaune, Côte de Nuits, Côte chalonnaise, Mâconnais). The data has been generated in different interactive situations in which communication about wine is the common denominator: wine teaching in wine school, wine tasting at fairs and wine presentations at wine Domains. We have identified two main categories of wine discourse production, a didactic discourse containing pedagogical features and a marketing discourse containing business rhetoric features. The marketing discourse itself includes a direct marketing discourse in sales situations and a promotional marketing discourse (= image promotion of the region) at wine fairs. Table 1 below gives an overview of the corpus and recordings at the moment as the corpus is an open one that will be completed in the future so as to cover the whole wine region:

Situation	Number of recordings	Length
Didactic	4	4hrs.33min.20 sec.
Direct marketing	3	2hrs9min.5sec.
Promotional marketing	5	1hr.33min.19sec.
	12	**8hrs.15min.44sec.**

Table 1: Overview of corpus and recordings

[3] OenoLex is at the moment restricted to the Burgundy region but can be extended to other wine producing region and can also include foreign languages on the basis of comparable corpora (Teubert 1996).

4. Recordings and digitalisation

4.1 Standardised audio sources

The first task was to generate and compile a formatted corpus of recorded speech data in order to establish the empirical base of the dictionary. Prior to data generation and acquisition, the interviewers received some training in sound recording techniques and used a TASCAM DR-07 MKI digital recorder. Data was formatted according to the technical research standards stated by the guides of the *Très Grande Infrastructure de Recherche* for Digital Humanities in France (Huma-Num 2015). Audio recordings were produced using uncompressed 24-bit/96 kHz WAV format.

4.2 Audio recording processing

We have used the free *Audacity* (2015) audio software to delete segments irrelevant to lexicographic data acquisition such as remarks made by the interviewer during the recordings. We have also tried to preserve the features of authentic, spontaneous discourse, including variation of speech intensity and pauses, and refrained from applying noise reduction, clipping, or signal amplification.

4.3 Transcribing and tagging

In line with the lexicographic concept, we have chosen to make use of an orthographic transcription methodology for the acquisition of the generated data in order to preserve the integrality of speech productions. We provide annotations for turn takings, overlaps, pauses, non-verbal productions, background sound details, etc. We use the Sonal free sofware (2015) to process audio flows and transcriptions. The software is designed to manage audio and video files and matching transcriptions, and provides a number of functionalities like transcription formatting, alignment, annotating, tagging, quality balancing, and import of transcriptions from other programs. The Sonal software includes easy-to-use specific tools for linguistic analysis (POS tagging, frequency lists, concordances, sorting out). It is also possible to isolate and select single sequences in order to perform clean transcriptions. Once done, we use the synchronisation module to align transcription and recorded speech data. The corpus is tagged using pre-defined thematic tags for the encoding of thematic content sequences such as ‘process’, ‘definition’, ‘appellation’, ‘technique’, ‘wine presentation’, ‘definition’, ‘question’, ‘wine-food agreement’, ‘sensoriality’, etc., which are used for the extraction of examples along with a range of formal segmentation and discourse strategic identification markers such as conventional segmentation expressions (*au premier nez/au deuxième nez*), negatives (*pas, pas que*), locatives (*on est sur*..), etc.

5. Extraction of examples: the case of negatives

Extraction is ruled by the recognition of knowledge-rich contexts used in

terminographic applications (Meyer, 2001). In the extracted example below, the lemmatised descriptor **citron** is negated and illustrates the case of a widespread didactic strategy in which the teacher anticipates erroneous interpretations from the students and states which descriptors are right and which are not. By doing so, the teacher performs a normalisation of the wine description independently of what students might interpret, the ultimate purpose being to assert and establish the quality profile of the wine. Examples of this type in which knowledge is confronted to non-knowledge [4] can be obtained by the extraction of "**ne pas"-** and "**pas que"**-marked negative contexts.

[*Au premier nez*, *on est sur* des fruits plus mûrs, on n'est **pas** sur le côté **citron**, agrume, mais sur le côté prune. *Au deuxième nez*, on évolue sur des arômes un peu miellés, on déduit même une sucrosité. Quand vous l'aérez, il est assez frais, c'est **pas** compoté, cuit, il y a même aussi un côté floral, il n'y a **pas que** le fruit. *On est sur* des fruits plus mûrs, peut-être la pêche, la prune.]

As illustrated, the semi-automatic extraction of (good) examples relies on a 3 level text- and data-oriented system: (i) thematic content descriptors to identify context relevant sequences (in actual oral interactions the term *citron* could also be used in a digression on fruits or on aroma from a very general point of view whereas a good example for OenoLex will always be related to the evaluation function of the interaction); (ii) interaction descriptors to link the data with the two poles of specialised knowledge and non-knowledge and (iii) linguistic markers ranging from discourse structuring elements to mere chunks (*i.e.* the special use of *être sur* in wine-tasting discourse). One key feature of the software used is to automatically build up sub-corpora on the basis of the first two types of descriptors that can then be linguistically and lexicographically processed.

6. Concluding remarks

We have tried to show how our semi-automated extraction method relies on a slightly different treatment of lemmas and examples. In OenoLex the main function of the lemmas (the wines and their associated descriptors) is not to serve as primary objects of lexicographical description but to provide direct access to the types of discursive strategies in which they are instantiated. This is achieved by 'discursive' lexicographic examples. Expert users can then access and retrieve information not on the meaning of the lemma, which is not an issue for the professionals, but on the discursive strategies revealed by systematically analysed, annotated, and indexed examples of wine tasting interactions. Our selection criteria are not guided by the ontologies of the domain but are entirely pragmatic and subsumed to the professional interactions of the domain. Examples are still addressed to the lemmas but are not simple elicitations. By doing so, we promote examples to the central functional component of what might

[4] In fact, to the extraction of knowledge-rich contexts described by Meyer (2001) should be added the extraction of 'knowledge poor' contexts as performed here.

be called ‘discourse dictionaries’ and hope to contribute to the development of a general theory of good lexicographic examples.

7. References

Books:

Gunnarsson, B. L.(2009). *Professional Discourse*. London: Continuum.

Book sections:

Gautier, L. & Leroyer, P. (2015). 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, in press.

Meyer, I. (2001). Extracting knowledge-rich contexts for terminography. In D. Bourigault et al. (eds.) *Recent Advances in Computational Terminology*. Amsterdam/Philadelphia: Benjamins, pp. 279–302.

Paper in conference proceedings:

Leroyer, P. & Valentina, H. & Maazaoui, H. & Chevalier, F. & Gautier, L. (2015). Faire déguster et parler de son vin pour le faire aimer et le vendre : quelques stratégies dans des interactions producteur-client. *Actes de International Wine Symposium of Toulouse*, *Université Toulouse – Jean Jaurès*.

Websites:

Audacity. Accessed at: http://www.audacity.fr (25 May 2015)

Guide méthodologiques pour le choix de formats numériques pérennes dans un contexte de données orales et visuelles. Accessed at http://www.huma-num.fr/ressources/guides (25 May 2015)

Guide des bonnes pratiques du numérique. Accessed at http://www.huma.num.fr/ressources/guides (25 May 2015)

Michelfeit, J. GDEX in Sketch Engine. ENeL 12 February 2015, WG3, Vienna, Accessed at : http://www.elexicography.eu/working-groups/working-group-3/wg3-workshops/automatic-extraction-of-good-dictionary-examples/ (25 May 2015)

Sonal®. Accessed at http://www.sonal-info.com (25 May 2015)

Journal articles:

Gautier, L. & Hohota, V. (2014). Construire et exploiter un corpus oral de situations de dégustation : l’exemple d’OenoLexBourgogne. *Studia Universitatis Babes-Bloyai*, *Philologia* 59/4, pp.157-173.

Leroyer. P. (2013). Proposals for the Design of Integrated Online Wine Industry Dictionaries. *Lexikos* 23, pp. 1-18.

Teubert, W. (1996). Comparable or Parallel Corpora? *IJL* 9/3, pp.238-264.

Dictionaries:

Bottlenotes. Wine glossary and encyclopedia. Accessed at : http://www.bottlenotes.com/winecyclopedia/glossary (13 May 2015)

Coutier, M. (2011). *Dictionnaire de la langue du vin.* Paris: CNRS Editions.