A descriptive algorithm for a wine tasting lexicon corpus

Margarita Goded Rambaud
Universidad Española de Educación a Distancia – UNED (España)

Resumen

Se pretende mostrar los avances en las pruebas de validez de un procedimiento de identificación computacional de los componentes que constituyen el significado de las expresiones en el restringido subdominio de las notas de cata de los vinos. El procedimiento consiste en un algoritmo de enlace que incluye un conjunto de componentes etiquetados. Dichos componentes van desde los no lingüísticos, con etiquetas para la "entrada perceptiva" y el "conocimiento del mundo", hasta los propiamente lingüísticos, tales como analizadores y definiciones de diccionario. Se utiliza la metodología Clashing Identification Procedure (CIP), que permite la reducción progresiva del corpus a un tamaño manejable. El interés de diseñar un sistema de etiquetado semántico reside en su contribución a la identificación de las expresiones metafóricas y sinestésicas que se usan frecuentemente en las notas de cata, y también a las tareas de desambiguación. En definitiva, se trata de mostrar cómo deducir computacionalmente la información relevante para la construcción de las metáforas en las que se basan las notas de cata y cómo un diseño de este tipo permite conectar conocimiento lingüístico y enciclopédico de una forma efectiva.

Palabras clave: Notas de cata de vinos. Etiquetado semántico. Ontologías. Algoritmo descriptivo. Gramáticas. Clashing Identification Procedure (CIP).

Abstract

The aim of this paper is to show advances in testing the validity of a procedure of computational identification of the components that make up the meaning of expressions in the restricted sub-domain of wine tasting notes (WTN). This takes the form of a proposed linking algorithm, which includes a number of tagging components. These components range from non linguistic ones, with taggers for *perceptual input* and *world knowledge* to traditionally linguistic ones, such as parsers or dictionary description. The proposed methodology is the Clashing Identification Procedure (CIP). The CIP is presented as a procedure, where different types of lexical descriptors produce a clashing when combined in a string of components that take the form of an algorithm. This leads to a subsequent

reduction of the corpus until a manageable size of it is reached. The interest of designing tagging components for a WTN corpus lies on its contribution to the identification of metaphoric and synaesthetic expressions that are frequently used in them and to other disambiguating tasks. That is, how to computationally deduce information relevant to the construction of metaphors, and how the contribution of such design will help connecting linguistic and encyclopaedic knowledge.

Keywords: Wine tasting notes. Semantic annotation. Ontologies. Descriptive algorithm. Grammars. Clashing Identification Procedure (CIP).

1. Introduction (*)

This paper is organized as follows: firstly, some characteristics of the lexical field where the corpus is obtained are shown. Secondly, the claim of the structural similarity of ontologies, grammars, and the proposed algorithm is discussed, where Langacker's ontological separation between objects and interactions is debated. Thirdly, the proposal for tagging annotation is discussed in a few examples. Finally, the results of the validation of the identification of the referent using the Clashing Identification Procedure (CIP) are presented.

2. Sensory background for wine tasting lexicons

The analysis of the characteristics of the wine tasting lexicon has been approached starting from the general procedure for sensory analysis. The conventional wine tasting procedure includes three stages in which the senses of sight, smell, and taste participate in this precise sequence.

Some characteristics define the sense of odour, which neurologically differentiate it from other senses. This differentiation is shown in various ways, and affects the idiosyncratic verbalization of this sensory experience.

Sensory description of visual perception adopts a type of lexicalization, which in some languages includes both colour nominalization and adjectivation. Colour adjectives include an extensive range of colour descriptors. Even if there are languages where basic referents for colours and their descriptors are the same, both Spanish and English have lexicalized them separately. Most languages have also gone through a similar abstraction process. That is, the sense of sight codifies colour separately from the referents having these colours. Both Spanish and English pairings

sangre/rojo, cielo/azul, nubes/blanco, negro/noche
and
blood/red, sky/blue, clouds/white, night/black

show the highly anthropocentric origin of the referents, and how referents and colour description are separately lexicalized.

Similarly, there are also terms which describe the abstraction of a tasting sensation and which have been lexicalized in the basic tasting terms *sweet*, *sour*, *salty*, *bitte*r (Spanish *dulce*, *ácido*, *salado*, *amargo*); again understood both as attributive adjectives or nouns. A separate lexicalization of referents as sources of sensory experience is shown in the following pairings:

azúcar/dulce, limón/ácido, sal/ salado, vinagre/amargo
and
sugar/sweet, yogurt/sour, salt/salty, vinegar/bitter

However, the generic terms in both languages for smell subdivide broadly into

smell: aroma-fragance/stench
olor: aroma-fragancia/hedor

The only kind of lexicalization that appears possible at this stage is a basic discrimination between pleasant/unpleasant odours. A basic discrimination/ differentiation between acceptance and rejection of an odour could possibly be related to survival and evolutionary constraints.

The research gathered by Morot, Brochet and Dubourdieu (2001) show a number of factors affecting the complexity of the smell sensory experience. They also explored verbal categorization in the sense of smell and explain that, in contrast with other sensory modalities, the fact that there are no specific terms – different from their respective sources – to designate odours confirms their hypothesis of the neurologically weak association between smell and language.

Morot et al. also suggest that the weak connection between odours and language is probably due to the brain lateralization, which takes place in the processing of odours and its lexicalization. They add that, while language processing takes place in the left side of the brain for most people, the processing of odours is right lateralized. They also affirm that the sense of smell is unlikely to provide enough information to generate sounded decisions in contrast with the information provided by other senses. As a result, the vast majority of odours take the name of the objects emitting these odours.

They found that the strong influence sight has over the other senses produces certain alterations, and they analyzed the empirical evidence they gathered in the smell phase of a blind taste performed by wine professionals. In their experiment, they showed how, when red wines were dyed with tasteless and decolouring chemicals, professional oenologists were unable to differentiate the reds from the whites. That is, they produced strong evidence of the dependence of smell on the information provided by other senses.

Odours then are described either using the term for the object that emits that particular odour or using metaphorical/metonymic descriptions. Wine tasting notes frequently use combinations of both.

Viberg (1984) proposes the following hierarchy for sense codification:

sight > hearing > touch > smell > taste

and demonstrates that a verb that has its basic meaning in a sensory modality to the left in the hierarchy may have an extended meaning covering some or all sensory modalities below in such a hierarchy.

She also studied the range of cognitive meanings of perception verbs, analyzing the use of *see*, and the fact that both *see* and *know* cover the same meaning and are, in fact, the same verb in various languages. The metaphoric use of *see* and *know* has been widely documented in the cognitive literature, allowing a new perspective in the interpretation of verbs of perception.

However, it was not until Sweetser's (1990) and Popova's (2003) works appeared that these metaphoric verbal extensions were considered systematic rather than anecdotic. Sweetser applies the Lakovian notion of embodiment to build the conceptual metaphor MIND-AS-BODY. As it is well known, this means that our understanding of the cognitive domain is based on the systematic correspondences between the domain of the body and the domain of the mind. She shows how there is a correspondence in verbs of perception between physical perception and states of the mind, and explains how, when using a traditional componential analysis in the semantic description of these verbs, there are no shared identifiable meaning components between the features defining *see* and *know*. The only systematic correspondence must be identified in our common experience of seeing and knowing.

As mentioned above, languages such as Spanish or English codify the so-called higher senses, separating referents and their lexicalization by means of nouns and adjectives. It is the case that the codification of sight in these two languages is lexicalized in a similar way. Both languages use the same term for the noun in the name of the colour and for the corresponding adjective, although in each language they both use different grammaticalization structures for adjectivation. Possibly because, as argued by Paradis (2005), nouns and adjectives can be based on the same types of content structure, but they are differently construed in all cases, which is why they are traditionally categorized as two different parts-of-speech in languages that make that distinction.

That is, since the same conceptual content can be construed as profiling either an entity or a relation, both nouns and adjectives can be used depending on the different level of abstraction required by the communicative interaction.

And this is why particular aspects of a wine tasting are made salient using a variety of construals under development. Highlighting either entities or relations can be achieved in a number of ways. Metonymyzation, abstraction, summary, and sequential scanning and profiling are all special cases of construals of salience

(ibidem) involved in the construction of any ontology, and particularly in the construction of a possible ontology of the subfield of wine tasting.

3. Structural similarity of ontologies, grammars and the proposed algorithm

Nirenburg and Raskin (2004) explain how building comprehensive computational linguistic applications involves making many theoretical and methodological choices, and how frequently developers are unaware of having made them. These authors also attribute this disconnection to the fact that both computational linguistics and natural language processing tend not to dwell on their foundations, and they neither create the resources and tools that might eventually help researchers and developers to view the space of theoretical and methodological possibilities available to them and to figure out the corollaries of their theoretical and methodological decisions. It seems that not only Nirenburg and Raskin but other computational linguists have also reached the conclusion that an ontological approach is needed in order to cope with the problems faced in the past years of the Natural Language Processing (NLP) research.

The theory-free application used here is based on a no-model linguistic approach, which originates from a number of facts. Firstly, the fact that, by contrast with other scientific disciplines, a unified linguistic model is not yet available. Although the generative trend, considered as a constellation of approaches, all of them within the chomskian paradigm, has reached some kind of internal consensus, it also frequently ignores other perspectives. Within the functional/cognitive paradigm, there is no shared body of knowledge either upon which build new hypothesis or to develop new research trends. As a result, linguistic theory has not reached a situation where a body of basic principles of analysis and methodological procedures could be agreed and shared as happens in other scientific disciplines, where such a shared body of descriptive elements, principles, axioms, and methods of analysis has long since been established. In the field of theoretical linguistics, there is no common body of knowledge upon which further efforts can be built and/or can be taken for granted.

Secondly, the fact that gathering and analyzing a vast amount of data is now technologically possible opens the possibility of statistical approaches that can be interpreted from a variety of linguistic angles. However, it is frequently the case that empirical statistic analyses are mainly used to validate certain pre-established theoretical stances rather than help advance key linguistic issues.

Statistical approaches have long since been exploited from a variety of perspectives, particularly from the computational linguistics angle. For example, information retrieval using web sources has been a common practice since the very beginning of the use of Fellbaum's (1998) WordNet.

For these reasons, the present analysis rejects theory alignment. This no-model approach comes together with the alternative proposal that the selection and configuration of tagging components to be used in a corpus encapsulates a grammar of the used language.

Emulating in a computer what the brain does, when simultaneously processing different types of information, can be formalized by means of a simple descriptive tool: an algorithm. It is claimed here that there is a structural similarity of the ontology of an area of knowledge or experience, the grammar of the language in which this knowledge or experience is expressed, and the proposed descriptive algorithm for the lexical pieces in a restricted corpus.

If these three concepts are analyzed in some detail, it can be observed that all three (an ontology, a descriptive algorithm, and a grammar) are theoretical constructs with a similar structure for an extremely simple reason: they share their descriptive components (entities and relations). These components of different kinds and categories can be sub-classified and used to tag a corpus.

Because an ontology represents an area of knowledge, a lexical field can be described in terms of the ontology of that particular field, and this description can take the form of a grammar. A grammar, in turn, can take the form of a descriptive algorithm. They all share the same basic structure:

f(x), a...n

where f(x) represents a relation of a kind (verbal, adjectival, or prepositional) and a, b, c..., n represent the entities linked by f(x). Obviously, the components making up the algorithm can be identified, and, as a result, a tagging procedure, amalgamating a number of components, is proposed here.

All definitions of ontologies, taken from both philosophy and computational sciences, include the reference to the identification of entities of all kinds (abstract or concrete) and the identification of relations among these entities.

In the most traditional and easily shareable of its definitions, the grammar of a language, in turn, includes the identification of a series of categories valid for that particular language (for example, adjectives, in languages that have this category, cases, etc.) and a number of rules of use of such categories.

Finally, a compilation of different definitions of the term *algorithm* results as follows: A set of instructions or rules that apply to a set of elements with a definite purpose in a particular sequence. For example, a cooking recipe is an algorithm that looks very much like this:

<ingredients + sequence of cooking instructions + purpose of obtaining an edible result>

For these reasons, it is claimed here that if a lexical field in a certain language is described, the ontology that identifies its categories and the relations operating

among them can be related to the grammatical categories operating in that particular language, which, in turn, can also be formalized by means of a series of instructions to be performed upon a series of components. This is why the identification of the components that make up the meaning of a lexical entry in a restricted lexico-semantic field is a first step in the creation of a descriptive algorithm that will include the whole set of meaning contributors labelled under a Greek letter in table I.

The sequence in which these components will be activated is still ongoing work, but the necessity of such a sequencing process is acknowledged by the insertion of a rule (R), which accounts for applicability or not of sequencing, as will be further explained below.

This is why the proposal of a series of labels for the components of a descriptive algorithm of a lexical entry is a conclusion that makes sense.

3.1. The issue of sequentiality

The organization of information requires a kind of hierarchy of concepts. Because of this, the meaning of a lexical entry being broken up in a series of taggeable components is also subject to an organized hierarchy. This hierarchy is related to the level of abstraction of each component. It seems natural to observe that the more abstract the component, the more language independent it is.

It is a well known fact that sequentiality or linearity is a crucial property for symbolic systems such as language, because certain logical and mathematical properties, such as transitivity depend on this characteristic.

Since all human information processing can be fed into the system either in a simultaneous or in a sequential mode, sequentiality operates differently in the input and output stages of language processing and elicitation. That is, when a piece of language is processed, both contextual and audial information is simultaneously processed. However, linguistic elicitation takes place in a sequential mode only. Sequentiality or linearity is a particular characteristic of all human languages, and this fact was originally identified by Saussure (1945) in the early years of the 20th century. Syntagmatic relations constitute a direct consequence of this particular specificity of human languages. Sequenciality then should take the form of a rule, such that for each lexical piece sequentiality is or is not applicable.

The agglutination of this information is quite simple and is only subject to one general specification or generic rule affecting sequentiality. Since ontologies are conceived of as hierarchisized conceptual descriptions, sequenciality operates in hierarchies, as well as in linear structures.

This is important from the computational point of view because if sequentiality is not applied at a certain point of the structure, it will rule out the activation of the content of various subsequent slots in the hierarchy as irrelevant.

As a result, the sequencing of the elements to be inserted and the sequencing of the instructions both determine the structure of the descriptive algorithm.

Ontologies, as linguistically and computationally related objects, are conceived of as hierarchized conceptual descriptions, where sequentiality is or is not applicable.

4. Semantic and ontological perspectives to be considered in the field

Although a broad cognitive approach along the line of Langacker (1991) is acknowledged as a pervasive influence for this work, this paper's theoretical background does not hold strong affiliations with any particular linguistic theory. Furthermore, no attempt has been made to try to fit this kind of tagging into a particular linguistic model. On the contrary, a mild claim is being posed in the sense that most existing grammars or linguistic models contribute to lexical representation of meaning in a variety of ways, and that these variations should be accounted for rather than obliterated.

However, Langacker's influence does not come without problems. He differentiates between objects and interactions quite clearly (ibidem, p. 16). Objects are instantiated in space, having spatial locations, are discrete, and are stable along time and space. In addition, objects are defined as conceptually autonomous, whereas interactions are defined as conceptually dependent in the sense that interactions do not exist independently of its participants.

Langacker develops his analysis introducing the term *entity*, which he defines as

> anything one might refer to for analytical purposes: objects, relationships, locations, sensations, points on a scale, distances, etc. Crucially, it is not required that an entity be discrete, individually recognized, or cognitively salient.

That is, after having differentiated between objects and interactions, both concepts become interchangeable. It seems as if the basic ontological differentiation between entities and relations should be avoided. As a result, this differentiation is invalidated since his definition for entities includes now both objects and relationships.

Further down, he explains that

> It is not the character of individual entities that is important, but rather the fact that they are interconnected and thereby constitute a region. (ibidem)

which is precisely why entities must be separate from relations. Or, alternatively, his previous differentiation between *objects* and *interactions* must be maintained.

But this is not the case, and the use of the term *entity*, including the concept of relationship is permeating further discussion and, to my view, misleading it. Particularly because, after listing what qualifies as entity (objects, relationships, locations, sensations, points on a scale, distances, etc.), he explains that any

expanse of material substance qualifies as an entity, thus highlighting objects and not relationships.

Since this basic differentiation operates in the organization of any ontology, it is unclear what would be the advantage of subsuming it in a more comprehensive concept or dispensing of it altogether.

Ontologies, either in their computational form or in their primary philosophy oriented configuration, are basically a set of entities that can be defined within a prefixed area of knowledge and a set of relations that can also be defined within the same area of knowledge. For the purpose of analyzing the particular semantic field of wine lexicon, the definition for ontology taken here includes the primary differentiation between entities and interactions in the line of Langaker's (ibidem, p. 4). In addition, if we understand ontologies as "a manageable instrument in the design of databases" (Paradis, 2005), these two concepts should be kept apart, firstly for the sake of clarity in the analysis, and secondly for both mental processing and computational reasons.

Whether the mind construes entities highlighting its relational, temporal, internal structure, such as, for instance, in the case of *property*, or highlighting its atemporal character, as Langacker proposes, is a another kind of analysis that can be attempted elsewhere. What is emphasized at this stage is only the fact that entities and the relations that can be identified among them constitutes a previous stage of analysis that the human mind is able to perform even before a subsequent linking of this differentiation to several parts of speech in the languages of the world is attempted.

In this differentiation, it is the abstract nature of second-order entities that is highlighted precisely because their relational nature is not. And precisely because of this, an ontology, as *methodologically providing us with manageable instrument in the design of a database*, must make a preliminary distinction between entities – of whatever kind – and relations as a starting premise.

5. Types of information included in the descriptive algorithm

Ontological semantics is related to the assumption that it is possible to reduce any natural language utterance to a formalized, language neutral representation (Moreno and Pérez, 2002). Since an ontology's main objective is to make explicit the conceptual organization of a particular field, they could be considered language independent. But, on the other hand, this knowledge is always lexicalized in particular languages of the world. Therefore, the meaning of a lexical entry can be formalized in a number of ways. In NLP, an ontology can be used to define lexical entries in ontological terms as it happens, for example, in Mikrocosmos' implementation of ontological semantics (Nirenburg and Raskin, 2004, p. 150).

The proposed descriptive algorithm has been designed to meet the requirements of a wine tasting note (WTN) and includes a number of taggers, which are described further down. The tagging procedure proposed in Goded Rambaud (2007) has been tried out using Robert Parker's WTNs (Wine Tasting Notes) as the main source of data. This set of WTNs has been compared with the BNC (British National Corpus).

Table I shows the different types of information that a lexical piece codifies in a WTN. These different types of information are described in such a way that they can take the form of a tagging component. All these different types of information should be codified accounting for the way in which they are captured and transmitted in natural language processing and in natural language use. It is claimed here that all these different types of information that have been labelled can be used to annotate a lexical entry.

The types of annotation used in this tagging procedure are of different kinds, ranging between clearly non-linguistic, such as [α] link to perceptual input, [ψ] world knowledge information or [Δ] link to a referent, to clearly and prototypically linguistic, such as [Ð] dictionary definition or [ω] conventional parsers.

Because a lexical entry allows for corpus treatment, it can be annotated. Presently, the study of this lexicalization is restricted only to Robert Parker's database. The main purpose of this field-based type of annotation is to capture the different linguistic and non linguistic components which, at different degrees of influence and at different levels of abstraction, contribute to creating the meaning of a WTN. While language independent tagging components are more closely linked to ontologies, language dependent annotation is related to corpora. I take the view that language dependent tagging components constitute a type of limited grammar that can take the form of a linking algorithm in order to be computationally processed.

Both types of components make up a set of elements, which constitute a proposed descriptive algorithm for the field. I make the claim that this descriptive algorithm includes all the specifications that a lexical entry needs to process the relevant information.

It is claimed here that the ontological requirements of particular fields of knowledge and/or experience, together with the characteristics of the corpora determine the appropriate selection of annotation. That is the tagging components that the annotations of the corpus require are field dependent.

The proposed descriptive algorithm is just an agglutination of components subject only to a general rule of organization. It is the selection of components and this general requirement that constitutes this limited grammar.

As in any ontology, two elements should be previously identified for this area of experience. One refers to the kind of entities present in the semantic field under study and the other to the type of relations that hold among these entities. An ontology describing the area of knowledge of food and wine is no exception.

Ways of accounting for sensory description making extensive use of synesthesia, metaphor, and metonymy, and ways of accounting for this in a proposed descriptive algorithm should be identified. This task is worthwhile to the extent that its findings can be extrapolated to other lexical fields but, most importantly, because some of these findings could be interpreted as contributions to the construction of a more general model of lexical representation.

5.1. Non linguistic annotation

The preliminary distinction between language dependent and language independent knowledge representations is again applied now. The fact that semantic descriptions are usually placed at a higher level of abstraction compared with syntactic descriptions does not mean that the former is language independent. Most semanticists from all affiliations agree that semantic description is lexically codified and, consequently, dependent on a specific language. As it is also widely accepted, semantic description is also related to syntax. For instance, the interface syntax-semantics operates syntagmatically at one key point that is where the argument structure fixes the type of contents to be acceptable in each argument slot. As a result, each verbal lexical predicate determines certain theta-roles configurations. This means that only a selection of components to be considered as descriptors in the proposed descriptive algorithm can be language independent. That is, [α] the link to perceptual input, [ψ] world knowledge information, [Δ] link to a referent, type of ontological construct [σ], type of logic construct [π]/[θ] and the configuration of schematic ontologies, such as part-whole [μ], degree [δ], frequency [φ], and boundness [β] are all language independent. All the rest are heavily dependent on the particular language used.

Non linguistic annotation, thus, includes taggers for [α], [ψ], [Δ], [σ], [π]/[θ], [μ], [δ], [φ], and [β].

The codification of sensory information [α] is related to the semantic inference of multimedia content. It requires a strong multidisciplinary collaboration to integrate the various sources of information. In the same line, encyclopaedic knowledge (ψ) is not always linguistically formalized, and therefore it is not easy to find out how the human mind is neurologically able to retrieve and relate the information needed to properly interpret language. Because of this, disambiguation programs in NLP are still insufficiently developed. Contributions, such as Smith and Brogaard (2003) add a notion of context to the resolution of vagueness and bind this notion into their ontological account but this is not enough

and, consequently, only the need to formalize this type of information is claimed at the present stage of development.

There are a few problems related to the codification of non-linguistic information. For instance, the type of cognitive information that codifies a certain perception of space, such that it allows the metaphoric description of a wine as *elegant* meaning *balanced*, is something still to be better understood. The processing of the type of socio-cultural information that allows us to interpret the description of a wine as *a whore* meaning *glyceric* is also poorly understood. The type of logic information, which includes a link to a referent [Δ] and to types of logical construct, is something well studied in semantics, logic, and the philosophy of language, but how the link is formalized so that the human processing takes place is still unclear.

All these types of information are accounted for by the human processor and humans use it to disambiguate the meaning of figurative language and metaphors.

5.2. Linguistic annotation

The access to linguist data, both in a manual and in a computational mode is an advantage that can be explored. In fact, it is an important source of empirical data for a wide number of research tasks.

In this case, three types of taggers are considered specifically linguistic: part of speech [λ], dictionary definition [Ð] and parsers/pragmatic, and discourse annotations all under the label [ω]. Although [λ] and [ω] are closely related, and under certain linguistic models, one can be subsumed under the other, information organized this way can be more profitably exploited.

5.2.1. The role of dictionary definition in monolingual dictionaries, the descriptive algorithm and CIP

It will be explained further down how the comparison of statistic occurrences between referential and non referential expressions in two corpora is used in relation to the role that [Δ] takes in relation with [Ð].

Dictionary information is the first type of specifically linguistic information that can be the subject of computational treatment in NLP. Dictionary definitions are a core element in the proposed descriptive algorithm. The inclusion of a dictionary definition component has processing advantages since it is a common facility already incorporated in PCs. It can be automatically linked to the corpus and this fact facilitates further subsequent manual annotations.

The importance of dictionary definitions in the design of linguistic models is not new. Coseriu (1983), Martín Mingorance (1998), Dik (1989), and Faber and Mairal (1999) among others, have recognized and used onomasiological definitions in their respective linguistic models.

A dictionary definition is a key element in the construction of an ontology, which is organized hierarchically according to the type of recognized categories. In addition, online web searches are closely dependent on field ontologies and make extensive use of online dictionary definitions.

Dictionary definitions in monolingual dictionaries have always been at the very core of semantic analysis, and the issue of how to produce an accurate definition of a word is tantamount to describing a concept. Because of this, dictionary definitions are closely related to the organization of knowledge in ontologies.

The connection between the definition of a word and its place in an ontology is not new either. It is as old as Aristotle. What is relatively new is the present possibility of linking dictionary definitions, corpora, and web searches thanks to the incorporation of all kinds of electronic dictionaries.

In this study, the component *dictionary definition* [Đ] is a key element in the design of both the descriptive algorithm and the clashing identification procedure (CIP), as explained below.

It has been made a central component in the design of the descriptive algorithm because it encapsulates different types of linguistic information. Once this information is computationally available, the possibility of manipulating it in order to contrast, compare, and/or equate it with other types of information, is a worth exploring task.

A computerized dictionary entry provides the type of information that can help identifying the target referent in a given processed text. The CIP allows for this since it is designed in such a way that contradictions among subcomponents of the different types of information encoded in the descriptive algorithm can be made explicit.

In addition, the position the different types of *definitions* take in a dictionary entry is indeed relevant. Electronic dictionaries are made after the manual compilation has taken place. That is, a human lexicographer has done the basic work of organizing the dictionary entry and including the different meanings of the word in it. This means there has been a human individual or collective process of decision making, and this previous process is worth exploiting for various reasons. Firstly, because it already implies the use of a pre-established criteria, and, secondly, because this criteria normally leads to a predetermined sequence for the presentation of the different types of meanings. That is, it carries a human reasoned prioritization.

The standard lexicographic procedure usually takes the first meaning of the word as referential. This first interpretation is normally followed by several others, which include figurative language interpretations, most frequently metaphoric and

other widely accepted contextual interpretations. Therefore, the position of the selected interpretation in a dictionary entry is relevant and can be used to make advances in the identification of the referent. Because of this, this type of layout can be used to help computational identification of these prefixed types of information.

As a result, if the position of a selected interpretation in a dictionary entry is considered to be relevant, this positioning can be used to discriminate referential and non referential interpretations, and can be exploited in the development of disambiguating programs or in the identification of figurative language.

5.2.2. Parsers

The range of available parsers is wide. Most provide only morpho-syntactic annotation; others include onto-semantic information as well. Because usually parsers have been developed as applications of the different linguistic models, their components are dependent on the mother linguistic theory. Since this paper does not claim to be faithful to any particular theory, any parser, which is robustly built, has its categories and relations clearly established, and accounts for some kind of selection restrictions, can be used here.

Since meaning is encoded at all levels of description (morphological, syntactic, and semantic), each in its own mode of capturing it, any ontology should include different types of information to be conflated in the proposed descriptive algorithm. At the level of morphological description, it should include part-of-speech basic information together with other collocational peculiarities of this particular lexical item. That is, once the basic part of speech classification has been completed, the morphosyntactic characterization should take the form of common parser, again complemented with collocational data. The main challenge, however, is to devise a system that allows for some kind of codification for those linguistic aspects, which highly affect meaning but which cannot be formalized along the same lines that logical or more directly referential lexical items do, such as happens with metaphoric, metonymic, and synesthetic linguistic resources.

Table I in the Appendix section shows how the description of each lexical item requires a number of separate semantic and cognitive entries arranged along a cline. This is represented by means of a dotted line to express the fact that there is no claim made of a separation between cognitive and semantic approaches in the descriptive tagging components, albeit the breakdown of them might suggest otherwise. The task to represent this conceptual overlapping, or this lack of clearly defined boundaries, which most probably will need the use of intervals, is left to mathematicians and computational experts.

A common parser providing morpho-syntactic information should not present processing problems in computational terms, and most grammars of different

affiliation have their own computational implementation. Pragmatic and discourse tagging is also linked to parsing.

The part-of-speech, dictionary definition and logical constructs have already been addressed in computational linguistics. It is the codification of the cognitive components that is more challenging and that still needs to be formalized in a straightforward way. And this is precisely what the restricted field of wine tasting description requires. How to account for these aspects and capture them in a particular tagger, specifically designed for this task, is one of the main challenges of the present work.

5.2.3. Ontological annotation

Ontological annotation (Δ) is placed close to logical annotation (π, θ) and to cognitive taggers accounting for configuration markers, such as part/whole (μ), degree (δ), frequency (φ) or boundness (β) applicable to different types of ontological constructs, such as entities or relations. This is because these taggers refer to what the term profiles in a particular context and in this particular corpus. This type of annotation includes type of category (ε) and type of ontological construct (σ) because the isolation of these aspects will help pinpoint the relevant profile of the lexical entry.

Following Lyons (1977) and Paradis (2005) the heading for [ε] selects a first, second or third order category. The type of ontological construct is related to simultaneous or subsequent assignment of the entity or relation annotation. It is placed next to the type of logical construct because of their obvious connections. In addition, these specifications are also linked to part-of-speech [λ] and are also connected to different kinds of profiling.

Ontological annotation [Δ] is placed close to logical annotation [π, θ] and to cognitive taggers accounting for configuration markers, such as part/whole [μ], degree [δ], frequency [φ] or boundness [β] applicable to different types of ontological constructs, such as entities or relations because these taggers refer to what the term profiles in a particular context and in this particular corpus.

Because of this, the connection between the preferred profiling and the selected type of logical construct has to be developed in relation with the hierarchy of components needed to construct the descriptive algorithm of a lexical entry.

6. The descriptive algorithm

This tagging procedure leads to the proposed descriptive algorithm:

[tagging components + general rule of sequentiality]
[[(α),(ψ),(Δ),(λ),(ω),(Ð),(ε),(σ),(π),(θ),(μ),(δ),(φ),(β)].R]

where the different types ofinformation provided in each slot contributes to the meaning of the lexical entry. In table I, a few examples of these different types of

information relevant for each example are marked. When there is no mark, it means that the type of information under this tag is not relevant or non existent.

7. The clashing identification procedure (CIP)

The CIP is related to a well-known procedure in syntactic analysis known as selection restrictions. It was developed within the generative paradigm very early in Chomsky's works (Chomsky, 1965) and has to do with how semantic limitations affect parsing. However, while selection restriction is based on present or positive evidence of contradictions or limitations in the co-occurrence of certain lexical entries in certain syntactic arrangements, CIP, on the other hand, is based on negative evidence of components analysis. Negative evidence is defined here as the relevance of the lack of information at a certain stage.

The procedure simply consists of assigning dictionary definition, parsing, and semantic applicability in order to identify preliminary clashings.

Firstly, the lexical entries that *cannot* be given a dictionary definition are, by default, the *sub-corpus* in which focus. Therefore, you either start searching for metaphors or other unconventional sources of ambiguity or to relate this to the identification of referential and non referential lexical items. That is, the lexical entries without this annotation (Ð) will be sub-corpus where other instructions are to be performed:

Σ = wine tasting notes corpus
Σ – Ð = manually taggeable corpus

Secondly, the <part-of-speech> [γ] component helps factor out a number of elements, leaving nouns and adjectives as the most likely head words supporting metaphor-like constructions. The stoplist facility is used to perform this type of filter.

Because configuration involving <part / whole > [μ], <frequency> [φ], <boundedness> [β], <degree> [δ], are more frequently linked to adjectives and adverbs, these subcomponents are to be linked to <perceptual input> [α], <part of speech > [γ] and referent (Δ), to identify clashes.

For example, the lexical entries for *wine*, *beauty* and *aggressive* can be found in sentences such as:

1. This is a beauty of a *wine*
2. This *beauty* will drink well over the next ten years
3. This wine is *aggressive*

These lexical entries take the following descriptive algorithm:

wine
{[(λ: noun), (Ð), (ε1), (θ), (δ), (β].R}

beauty
{[(λ: noun), (Ð: ø), (θ), (δ).].R}

aggressive

{[(λ: adjective), (Ð), (π),(μ),(δ), (β)].R}

It can be observed that the component type of ontological construct [σ] for *wine* or *beauty* will be <entity>, whereas for *aggressive*, it will be <relation>.

The components Dictionary Entry [Ð] and Link to Referent [Δ] for *aggressive* and *beauty* will clash with the component link to perceptual input [α]. It will be ruled out and not be part of the descriptive algorithm of the lexical entry.

Wine and *beauty* both share [σ] type of ontological construct, but they will clash in type of category (1st vs. 3rd order entity).

In the case of *wine*/*beauty*, the entry for *wine* (central in any wine tasting notes corpus) and the entry for *beauty* in (ii) both clash. These entries are both nouns but, whereas a dictionary entry is applicable in the case of *wine*, in the case of *beauty*, the metaphorical description is not compatible with a dictionary entry, and therefore, a *clash* can be identified. They are both different types of entities and in both cases; a description of qualia is applicable.

The case of *beauty aggressive* also shows how a clashing operates. Again, the component type of ontological construct [σ] for *beauty* will be <entity>, whereas for *aggressive* it will be <relation>.

The component dictionary entry (Ð) for *aggressive* and *beauty* will clash with the component link to perceptual input [α] and Link to Referent [Δ], and it will be ruled out and not included in the descriptive algorithm. This will help reduce the general corpus (∑) to a manually taggeable one (∑ - Ð).

8. Verification of the CIP

The purpose of the experiment, whose results can be seen in table II, is to obtain statistic validation of lexical entries used in referential and non referential collocations. By doing this, it is possible to see what company a selection of lexical entries keep and how they collocate. Their collocation is used to identify a clashing or contrast of components. The first two components of the descriptive algorithm validated are the dictionary entry component [Ð] and the link to referent [Δ].

The computational identification of clashings among tagging components is linked to the identification of non referential expressions (Nazar, 2008). To do this, statistic analysis has been used to measure how a number of referential and non referential lexical entries, randomly selected from the Robert Parker corpus, collocate in this corpus (WTN corpus) and how they collocate in the BNC.

The procedure used to obtain computational validation of the clashing identification started with a manual identification of a selection of metaphoric expressions in a number of WTN corpus. This was followed by an identification

of their collocations in the WTN corpus using WSMTH tools which produced a preliminary list. After that, occurrence ratios in the Robert Parker corpus were calculated. Then an analysis of the correspondence of referential and non-referential expressions was performed to identify the correlations with the highest ratios of occurrence.

A similar identification of collocations of the same lexical entries in the BNC was performed to produce another preliminary list and to calculate ratios. Again, an analysis of correspondences was performed to see if the highest ratios correlated with referential or with non-referential expressions.

Finally, both frequencies were compared to check the initial hypothesis and it was confirmed in the sense that collocations of the selected items in the BNC tend to be more referential than the collocations of the same items in the WTN corpus.

9. Comparative analysis of collocations in both corpora

Table 2 shows the results of statistic collocations for the selected entries.

It can be observed how the entries *aggressive*, *angular*, *attack*, *balance*, and *beauty* collocate and what they combine with and in which frequency. This statistic analysis shows that *aggressive* in the Robert Parker corpus collocates with *tannin*, *oak*, and *acidity* whereas in the BNC, *aggressive* collocates with *behaviour*, *marketing*, and *stance*.

It also shows how *angular* collocates with *wine*, *finish* and *austere* in the Robert Parker corpus while it collocates with *face*, *shape*, and *fragment* in the BNC. Similarly, *attack* collocates with *sweet*, *fruit*, and *palate* in the WTN corpus but it collocates with *bomb*, *aircraft*, and *victim* in the BNC.

While *balance* collocates with *fruit*, *wine*, and *purity* in the WTN corpus, it collocates with *account*, *payment*, and *effect* in the BNC. And finally, if *beauty* collocates with *drink*, *finish*, and *bodied* in the WTN corpus, it is found in the company of *spot*, *salon*, and *products* in the BNC.

That is, the most frequent collocation in the Robert Parker corpus is a non-referential collocation, which is no news at all. However, what is highly relevant is that the most frequent collocation of the same word in the BNC is always referential.

In addition, no cases found in the BNC are statistically relevant, where the lexical entries studied collocate with other lexical entries in a non-referential combination which, again, contrasts with the fact that there are no statistically relevant cases found in the WTN corpus, where the same entries collocate with other lexical entries in a referential combination.

10. Conclusions

The lexical field that codifies the sensory experiences of wine drinking has been selected because it constitutes a very adequate testing ground for trying out different alternatives in lexical codification. The proposed lexical codification takes the form of a descriptive algorithm, which amalgamates various kinds of semantic and cognitive information.

The structural equivalence of ontology of a wine field, the grammar of the language used and the descriptive algorithm proposed for each lexical piece has been discussed.

In this paper, I have tried to show how the description of each subfield calls for different combination of representational tools. On the one hand, the most clearly referential terms in the field could have done with a similar referential description componentially based. On the other hand, sensory description requires not only highly synesthetic adjectivation, but most importantly, the peculiarities of the adjectivation used must be approached from a cognitive perspective, which could account, among others, for boundary problems. In addition, the highly metaphoric type of description in wine tasting requires an approach, which should be able to combine various perpectives.

The proposed descriptors take the form of an amalagamation of elements, which can be attached to each lexical item or construction. The list of taggeable descriptors is organized along two main broad categories: linguistic and non linguistic, which is subsequently divided in sections along a line. Each lexical piece or construction is specified for semantic description under part of speech, dictionary definition and type of ontological category. Under the cognitive description the part/whole, degree/frequency, and boundness should be specified mainly as leading to internal relations or functions. Somewhere in between, both kinds of logical constructs – predicate and qualia structure – are inserted. A general rule accounting for sequentiality will be applicable to all descriptors. Further specification will set the hierarchical relations among descriptors. It is claimed here that this simple descriptive algorithm, ontologically based, and strongly field dependent, is basically a nuclear grammar to be computationally implemented.

In addition, a statistical procedure to identify when a link to a referent [Δ] and a dictionary entry [Đ] are incompatible, components have been developed. This shows how the CIP is a possible valid method that can help identify contradictions among components of the descriptive algorithm. How this procedure is used in several applications, such as metaphor identification or the improvement of disambiguating programs is now under development.

Acknowledgements

I am indebted to Alfredo Poves Luelmo, from the Universidad Complutense de Madrid, for his help comparing corpora and with the calculation of the figures in table II.

Notes

(1) Collins Cobuild Dictionary.

Editor's note

(*) This paper is an extended version of the paper presented at Ibersid 2007 and published in *Ibersid: Revista de Sistemas de Información y Documentación*. (2007) 313-321.

References

Bateman, John A. (2007). Linguistic interaction and ontological mediation. // Schalley, Andrea C.; Zaefferer, Dietmar (eds.). Ontolinguistics: How ontological status shapes the linguistic coding of concepts. Berlin / New York: Mouton de Gruyter, 2007. 115-144.

Chomsky, N. (1965). Aspects of the Theory of Syntax. Cambridge, Mass.: MIT, 1965.

Coseriu, Eugenio (1983). Introducción a la lingüística. México: UNAM / Instituto de Investigaciones Filológicas / Publicaciones del Centro de Lingüística, 1983.

Dik, S. C. (1989). The theory of functional grammar. Dordrecht / Providence: Foris, 1989.

Eco, U., M. Santambrogio, P. Violi (1988), eds. Meaning and Mental representations. Indiana University Press. Bloomington and Indianapolis.

Faber, Pamela; Mairal Usón, Ricardo (1997). Definitional analysis in the functional-lexematic lexicographic model. // Alfinge: Revista de Filología. 9 (1997) 217-232.

Faber, P. B.; Mairal, R. (1999). Constructing a lexicon of the English verbs. Berlin: Mouton de Gruyter, 1999.

Fellbaum, Christiane (ed.) (1998). WordNet. An electronic database. Cambridge, MA: MIT, 1998.

Goded Rambaud, Margarita (2007). A descriptive algorithm for a wine tasting lexicon corpus. // Ibersid: Revista de Sistemas de Información y Documentación. (2007) 313-321.

Herdenstam, Anders (2004). Experience of an aesthetic sensation: Wine tasters in the field between art and science. Licentiate thesis. Stockholm: Royal Institute of Technology. Department of Economics and Management, 2004.

Herdenstam, Anders P. F.; Hammaren, Maria; Ahlström, Richard; Wiktorsson, Per-Axel. The proffesional language of wine: Perception, training and dialogue. // Journal of Wine Research. 20:1 (2009) 53-84.

Johnson-Laird, P. N. (1988). How Is Meaning Mentally Represented? // Eco, U.; M. Santambrogio, M.; Violi, P. (eds.). Meaning and Mental Representations. Bloomington; Indianapolis: Indiana University Press. 99-117.

Lakoff, G.; Thompson, M. (1980). Metaphors we live by. Chicago, IL: The University of Chicago Press, 1980.

Langacker, R. (1991). Foundations of Cognitive Grammar. Vol II. Stanford, CA: Stanford University Press, 1991.

Lehrer, Adrienne J. (1983). Wine and conversation. Bloomington, IN: Indiana University Press, 1983.

Lehrer, Adrienne J. (1985). Is semantics perception driven or network-driven? // Australian Journal of Linguistics. 5 (1985) 197-209.

Lehrer, Adrienne; Kittay, Eva Feder (1992). Frames, fields and contrasts. Hillsdale, NJ: Lawrence Earlbaum Associates, 1992.

Lyons, J. (1995). Linguistic Semantics. Cambridge: Cambridge University Press, 1995.

Lyons, J. (1977). Semantics. Cambridge: Cambridge University Press, 1997. 2 vol.

Mairal, R.; Pérez Quintero, M. J. (2000). New perspectives on argument structure in Funcional Grammar. Berlin: Mouton de Gruyter, 2000.

Martín Mingorance, Leocadio (1990). Functional grammar and lexematics in Lexicography in Tomaszczyk, J.; Lewandowska-Tomasxczyk, B. (eds). Meaning and Lexicography. Amsterdam: John Benjamins, 1990.

Martín Mingorance, Leocadio (1998). El modelo lexemático-funcional: el legado lingüístico de Leocadio Martín Mingorance. Granada: Universidad de Granada, 1998.

Moreno, A.; Pérez, C. (2000). Ontological semantics and lexical templates. // Mairal, R.; Pérez Quintero, M. J. New perspectives on argument structure in funcional grammar. Berlin: Mouton de Gruyter 2000.

Morot, G.; Brochet, F.; Dubourdieu, D. (2001). The colour of odors. // Brain and Language. 79 (2001) 309-320.

Nazar, R. (2008). Diferencias cuantitativas entre referencia y sentido. // Actas del XXVI Congreso de AESLA (Universidad de Almería, del 3 al 5 de abril de 2008). http://www.upf.edu/pdi/iula/rogelio.nazar/SinnUndBedeutungFeb27.pdf (2008-10-10).

Nickles, Matthias; Pease, Adam; Schalley, Andrea C.; Zaefferer, Dietmar (2007). Ontologies across disciplines. // Schalley, Andrea C.; Zaefferer, Dietmar (eds.). Ontolinguistics: How ontological status shapes the linguistic coding of concepts. Berlin / New York: Mouton de Gruyter, 2007. 23-67.

Nirenburg, S.; Raskin, V. (2004). Ontological Semantics. Cambridge, MA: MIT Press, 2004.

Paradis, C. (2004). Where does metonymy stops? Senses, facets and active zones. // Metaphor and Symbol. 19:4 (2004) 245-264.

Paradis, C. (2005). Ontologies and construals in lexical semantics. // Axiomathes. 15 (2005) 541-573.

Paradis, C. (2007). Configurations, construals and change expressions of degree. Submitted to English Language and Linguistics.

Parker, Robert (2007). http://www.erobertparker.com/ (2008-10-10).

Popova, Y. (2003). The fool sees with his nose: metaphoric mappings in the sense of smell in Patrick Süskind's Perfume. // Language and Literature. 12:2 (2003) 135-151.

Saussure, F. (1945). Curso de lingüística general. Buenos Aires: Losada, 1945.

Schalley, Andrea C.; Zaefferer, Dietmar (eds.) (2007a). Ontolinguistics: How ontological status shapes the linguistic coding of concepts. Berlin / New York: Mouton de Gruyter, 2007.

Schalley, Andrea C.; Zaefferer, Dietmar (2007b). Ontolinguistics: An outline. // Schalley, Andrea C.; Zaefferer, Dietmar (eds.). Ontolinguistics: How ontological status shapes the linguistic coding of concepts. Berlin / New York: Mouton de Gruyter, 2007. 3-22.

Scott, Mike (2008). WordSmith Tools 5.0. Oxford: Oxford University Press, 2008.

Smith, Barry; Brogaard, Berit (2003). A unified theory of truth and reference. // Logique et Analyse. 43:169-170 (2003) 49-93.

Sweetser, E. (1990). From etymology to pragmatics: metaphorical and cultural aspects of semantic structure. Cambridge: Cambridge University Press ("Cambridge Studies in Linguistics"), 1990.

Viberg, A. (1984). The verbs of perception: a typological study. // Butterworth, B.; Comrie, B.; Osten, D. (eds.). Explanations for Language Universals. 123-162. Berlin: Mouton, 1984.

Wine guides

Peñín, José (2006-2007). Guía Peñín de los vinos de España. Madrid: Pi &Erre.

Parker, R.(2007). *Wine Advocates*. http://www.erobertparker.com/

Sofware used

Mike Scott's WordSmith Tools.

Appendix

WTN corpus Σ	**<non linguistic tagging> .... semantic tagging ..... cognitive tagging .....>**													
Lexical item or construction	Link to **perceptual input**	**World knowledge** information	Link to **referent**	**Part of speech**	**Parsing, P & D annotation**	**Dictionary entry**	Type of **category**	Type of **ontological construct**	Type of **logical construct**		configuration			
											Schematic ontologies			
	α	ψ	Δ	γ	Ω	The role of dictionary definition in the descriptive **algorithm and in the CIP**	ε	σ	Predicate	Qualia structure	Part / whole	Degree	Frequency	Boundedness
									π		μ	δ	φ	β
wine				noun		An alcoholic drink, which is made from grapes and is usually either red or white (1)	1st order entity	entity		+	+			+
beauty				noun		Ø	3rd order entity	entity		+		+		
aggressive						Ø		relation	+			+		

Table I. Tagging proposal.

Wine Tasting Corpus (WTC)						British National Corpus (BNC)					
Collocates						Collocates					
		Total	*C/W (%)*	*Total WTC*	*as C (%)*			*Total*	*C/W (%)*	*TotalBNC*	*as C (%)*
aggressive	tannin	140	34,23	18356	0,76	**aggressive**	behaviour	81	4,10	12033	0,67
	oak	82	20,05	16543	0,50		marketing	22	1,11	5220	0,42
	acidity	56	13,69	13411	0,42		stance	21	1,06	1681	1,25
		Total	*C/W (%)*	*Total WTC*	*as C (%)*			*Total*	*C/W (%)*	*Total BNC*	*as C (%)*
angular	wine	81	21,26	81039	0,10	**angular**	face	10	0,51	33463	0,03
	finish	74	19,42	22246	0,33		shape	9	0,46	6780	0,13
	austere	67	17,59	2012	3,33		fragment	3	0,15	942	0,32
		Total	*C/W (%)*	*Total WTC*	*as C (%)*			*Total*	*C/W (%)*	*Total BNC*	*as C (%)*
attack	sweet	405	29,76	20645	1,96	**attack**	bomb	17	0,18	3126	0,54
	fruit	377	27,70	40877	0,92		aircraft	16	0,17	6232	0,26
	palate	254	18,66	10244	2,48		victim	16	0,17	3461	0,46
		Total	*C/W (%)*	*Total WTC*	*as C (%)*			*Total*	*C/W (%)*	*Total BNC*	*as C (%)*
balance	fruit	81	21,26	40877	0,20	**balance**	account	132	2,82	16342	0,81
	wine	74	19,42	81039	0,09		payment	59	1,26	5581	1,06
	purity	67	17,59	6227	1,08		effect	31	0,66	23093	0,13
		Total	*C/W (%)*	*Total WTC*	*as C (%)*			*Total*	*C/W (%)*	*TotalBNC*	*as C (%)*
beauty	drink	328	29,63	28248	1,16	**beauty**	spot(s)	95	2,31	4957	1,92
	finish	140	12,65	22246	0,63		salon	39	0,95	666	5,86
	bodied	108	9,76	40439	0,27		products	35	0,85	11311	0,31

Table II. Referential and non-referential collocates.

Recibido: 2009-05-30. Revisado: 2009-08-29. Aceptado: 2009-09-09