How Terms Meet in Small-World Lexical Networks: The Case of Chemistry Terminology

Francesca Ingrosso, Alain Polguère

How Terms Meet in Small-World Lexical Networks: The Case of Chemistry Terminology

Francesca Ingrosso
SRSMC UMR 7565,
CNRS-Université de Lorraine
Boulevard des Aiguillettes, BP 70239
54506 Vandœuvre-lès-Nancy Cedex, France
francesca.ingrosso@univ-lorraine.fr

Alain Polguère
ATILF UMR 7118,
CNRS-Université de Lorraine
44 av. de la Libération, BP 30687
54063 Nancy Cedex, France
alain.polguere@univ-lorraine.fr

Abstract

We present a new type of terminological model based on formal network structures called *lexical systems*. Those are non-hierarchical lexical graphs where the bulk of lexical relations is formally encoded by means of Meaning-Text lexical functions. This paper describes how this approach to lexical structuring can be applied to the modeling of terminologies, more specifically, to the French and English terminology of chemistry. The first section explains the importance of terminology in chemistry and introduces the aim of our project. Section 2 is a brief presentation of formal characteristics of lexical systems. Section 3 illustrates the type of terminological descriptions we are implementing with the specific case of the chemical term *catalysis*.

1 Structuring the Lexicon of Chemistry

1.1 Key Role of Terminology in Chemistry

Terminology plays a key role in chemistry research. For instance, chemical terms, by their very morphological structure, are closely related to the behavior and properties of substances they designate. As noted by R. Hoffmann[1] and P. Lazlo (1991), the knowledge of the name of a chemical compound, that strictly reflects the compound structure, gives the chemist the "control" over the molecule. Additionally, the terminology of chemistry is extremely vast and fluctuant. The importance of using a proper terminology in chemistry has lead to the creation, in 1919, of the IUPAC: International Union for Pure and Applied Chemistry. IUPAC elaborates rules for the nomenclature of molecules, in order to avoid definitional ambiguities and ensure harmonization of terminological proposals when new molecules are discovered. It has made available on-line for chemists the so-called *Gold Book* (McNaught and Wilkinson, 1997): an extensive dictionary-like description of English chemistry terms.

In spite of such efforts, the terminology of chemistry is loosely normalized. There is also a lack of multilingual perspective as most scientific papers are written in English, which can lead to serious problems in the context of the teaching of this discipline in schools and universities.

1.2 Terminological Networking

In the on-line IUPAC *Gold Book*, term descriptions are eminently relational, as illustrated by the entry for the nominal term *bond* below.[2]

(1) **bond**
There is a chemical bond between two atoms or groups of atoms in the case that the forces acting between them are such as to lead to the formation of an aggregate with sufficient stability to make it convenient for the chemist to consider it as an independent 'molecular species'.
See also: agostic, coordination, hydrogen bond, multi-centre bond

The definition (*There is a chemical bond ...*) in this description is doing its job of establishing connections between *bond* and related terms such as *force*, *aggregate*, etc. It is however an unstructured and non-formalized model for such connections. A greater applicative potential could

[1] Nobel Prize in Chemistry 1981.

[2] http://goldbook.iupac.org/B00697.html

be achieved by explicitly encoding the linguistic structure of a term definition. If such structure mirrors the logical organization of concepts in the corresponding scientific domain, the lexical definition can be an efficient tool for the understanding of scientific texts, for scientific writing and for teaching chemistry.

Additionally, as illustrated by terminological pointers at the end of (1) – *agostic*, *coordination*, *hydrogen bond*, *multi-centre bond* – the mastering of a chemistry term and of the corresponding notion depends on the ability to position this term within the network of other terms that gravitate in its semantic space.

Polysemy is another acute problem in the terminology of chemistry, that is generally ignored in existing resources. Polysemy manifests itself in two ways.

A) It can occur within the terminology itself, when a single form is used to denote different terminological notions – for instance, *to catalyze* as '[for a substance] to cause a certain type of chemical reaction' in (2) *vs.* '[for a chemist] to make this reaction take place' in (3):[3]

(2) *These fiber catalysts can efficiently* ***catalyze*** *the Knoevenagel condensation of benzaldehyde and ethyl cyanoacetate in water (yields: 95-98%).*

(3) *These Ta2O5-T samples were characterized by TG / DTA, XPS, nitrogen adsorption, XRD, and UV-Raman, and were employed to* ***catalyze*** *the gas-phase dehydration of glycerol (GL) to produce acrolein (AC) at around 315 degrees C.*

B) Polysemy can also spread over both chemistry terminology (2)-(3) and general language (4).[4]

(4) *Cities are always building new stadiums with the justification that they'll* ***catalyze*** *the local economy.*

All these observations show that it is necessary to organize the terminology of chemistry according to rigorous theoretical and descriptive principles.

The project we are presenting has a very practical aim: the design and construction of a terminological database in chemistry, for both the English and French languages. It also has theoretical implications as it explores a new approach to the structuring of terminologies based on non-hierarchical graph structures (see lexical systems, section 2 below), where each term is an element in a global lexical network in which it is related to the rest of the domain terminology, as well as to general language lexicon, by means of Meaning-Text lexical functions (Mel'čuk, 1996). Lexical functions have already proved to be an efficient tool to model relations between terms (L'Homme, 2002). In our project, however, the recourse to lexical functions is embedded within a formal proposal for the graph structuring of lexical knowledge – lexical systems – that we believe is particularly suited to account for the interaction between terminologies of different domains – e.g., chemistry terms used in physics – as well as between "purely" terminological units and units that belong to the general language – e.g. *water* as a type of molecule and *water* as a substance.

2 Terminologies as Lexical Systems

The terminological models we are elaborating are grafted on two general language lexical resources: the *English* and *French Lexical Networks* (Gader et al., 2014; Lux-Pogodalla and Polguère, 2011), respectively en-LN and fr-LN.

The design of the en- and fr-LNs is based on a new type of lexical model called *lexical system* (Polguère, 2014). From a formal point of view, a lexical system is a graph whose vertices are lexical units of the lexicon under description and whose edges are lexical relations of essentially two types:

1. semantic relations – (*chemical*) *bond* is linked to *to bond*, *interaction*, *compound* . . . ;

2. combinatorial relations – (*chemical*) *bond* combines with *covalent*, *ionic* . . .

Both types of relations are modeled by means of lexical functions (section 1.2 above): paradigmatic lexical functions in the first case and syntagmatic lexical functions in the second case. Though lexical functions provide the bulk of graph structuring in lexical systems, other types of relations are also implemented. For instance,

[3] Chemistry examples are borrowed from *Web of Science* (`http://webofscience.com/`).

[4] *New York Times*, COCA corpus (`http://corpus.byu.edu/coca/`).

semantic embedding is implemented via links weaved within lexical definitions: cf. the definition of CATALYSIS **I.1**, section 3.3 below, that formally links this term to two semantically embedded terms: REACTION **1** and GIBBS ENERGY.

Such graphs belong to the family of so-called *small-world networks* (Watts and Strogatz, 1998) and their topological properties allow for the automatic identification of semantic spaces through clusterization. Figure 1 illustrates the semantic space of BOND$_N$ **I.2** – the chemistry sense of the noun BOND$_N$ in the current version of the en-LN[5].

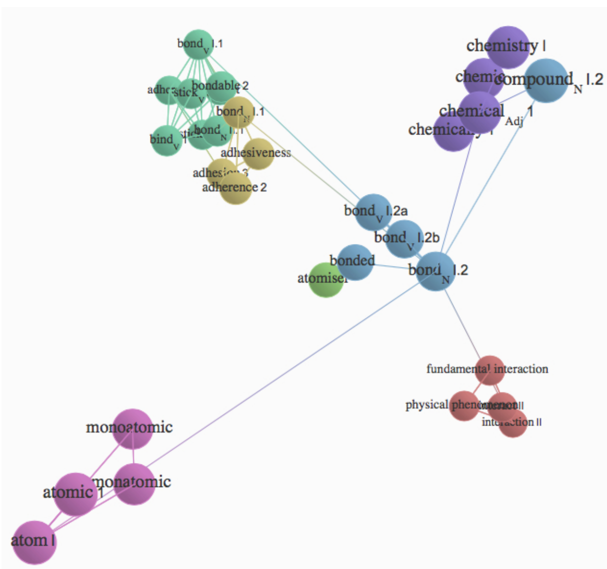


Figure 1: Semantic space of BOND$_N$ **I.2**.

Beside being computer-tractable structures, lexical systems are equivalent to “virtual dictionaries”, as all properties of lexical units are encapsulated in graph vertices – lexical definitions, grammatical information, citations from corpora (i.e. contexts), etc. Thanks to a specially designed lexical graph editor, it is possible to build a lexical model and, thus, a terminology, by methodically weaving lexical systems (Polguère, 2014). In the specific case of the work presented here, we are describing the terminology of chemistry by weaving the terminological network of this discipline directly on top of the general language en- and fr-LNs. This will allow not only for the proper connection of terminologies in both language, but also for the “interpretation” of chemical terms relative to the non-specialized lexical stock, with which these terms naturally interact in standard research activity as well as in scientific texts.

3 Example: The Case of CATALYSIS I.1

We now illustrate our approach with the en-LN description of the noun CATALYSIS. It possesses the same polysemic structuring as the corresponding verb CATALYZE – see (2), (3) and (4), section 1.2. Therefore, two chemistry senses have to be distinguished within the nominal vocable:

- CATALYSIS **I.1** [*The catalysis occurred via the formation of a chloromethylated triflate complex, and electrophilic addition to an aromatic hydrocarbon.*]
- CATALYSIS **I.2** [*Were they doing catalysis, and if so, how did they recover the catalyst?*]

We will focus on the term CATALYSIS **I.1**, which is the nominal counterpart of the basic sense of the verb exemplified in (2).

3.1 From Lexical Graph to Article-View

When weaving lexical systems with the tailor-made graph editor named *Dicet* (Gader et al., 2012), lexicographers are provided with a textual rendering of lexical information: the *article-view* of the headword. We present the article-view of CATALYSIS **I.1** in Figure 2 below.

It is essential to note that the article-view is only the textual display of fundamentally relational information encoded in the lexical network. For instance, what appears as:

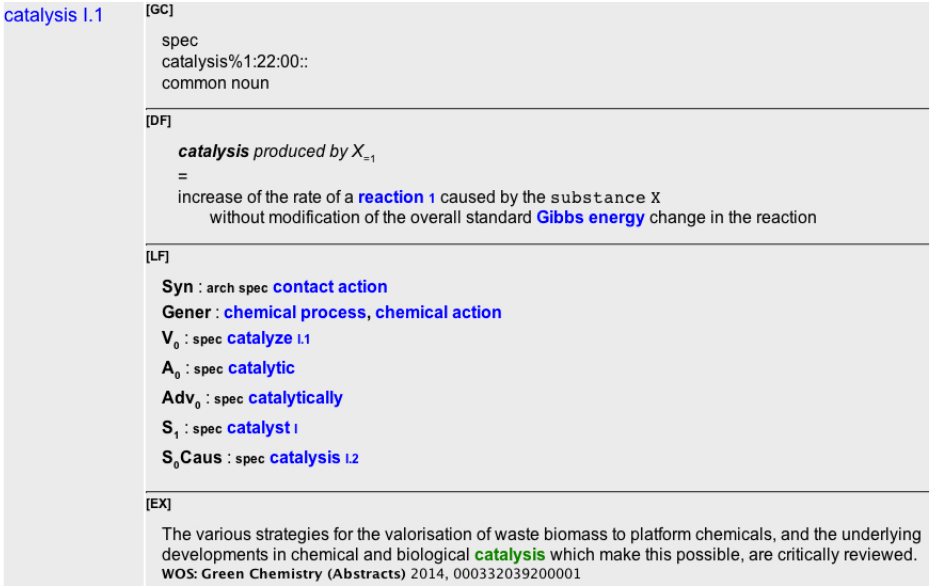
S_1 : **spec** catalyst **I**

in Figure 2 is generated (i) from an **S_1** lexical function link (typical name for the 1st actant of the headword) holding between CATALYSIS **I.1** and CATALYST **I** and (ii) from a grammatical characteristic link connecting this latter unit to the linguistic usage note “spec”, that characterizes CATALYST **I** as being a term.[6]

In order to truly apprehend the formal nature of the lexical model in which terminologies are embedded, it is therefore necessary to distance oneself from textual article-views and focus on the

[5] Graph visualizations are based on the Tmuse algorithm (Chudy et al., 2013) and are generated with tools provided by Kodex.Lab (`http://kodexlab.com`).

[6] Even citations – cf. the [EX](ample) zone in Figure 2 – are implemented as connections between individual citations and lexical units they contain.


catalysis I.1

[GC]

spec
catalysis%1:22:00::
common noun

[DF]

catalysis *produced by* $X_{=1}$
=
increase of the rate of a **reaction 1** caused by the `substance X`
without modification of the overall standard **Gibbs energy** change in the reaction

[LF]

Syn : arch spec **contact action**
Gener : **chemical process**, **chemical action**
$\mathbf{V_0}$: spec **catalyze I.1**
$\mathbf{A_0}$: spec **catalytic**
$\mathbf{Adv_0}$: spec **catalytically**
$\mathbf{S_1}$: spec **catalyst I**
$\mathbf{S_0Caus}$: spec **catalysis I.2**

[EX]

The various strategies for the valorisation of waste biomass to platform chemicals, and the underlying developments in chemical and biological **catalysis** which make this possible, are critically reviewed.
WOS: Green Chemistry (Abstracts) 2014, 000332039200001

Figure 2: Current en-LN article-view for CATALYSIS **I.1**.

core structuring element of our model: the multidimensional system of lexical function relations that connects lexical units.[7]

3.2 Web of Lexical Function Relations

The structurally most relevant information in Figure 2 appears in the lexical function zone `[LF]`. It corresponds to the set of paradigmatic links that originate from the CATALYSIS **I.1** headword and connect it the rest of the lexical system. (At present, no syntagmatic link has been encoded for this specific term.) It is this information, together with incoming lexical function links, that position the term in the global structure of the en-LN and defines its semantic space.

In our terminology, the semantic space of a lexical unit such as CATALYSIS **I.1** is much more than just the subgraph constituted of all outgoing and incoming lexical function links. It is the topologically significant cluster of semantically-related nodes that gravitate around CATALYSIS **I.1**, as illustrated in Figure 3.

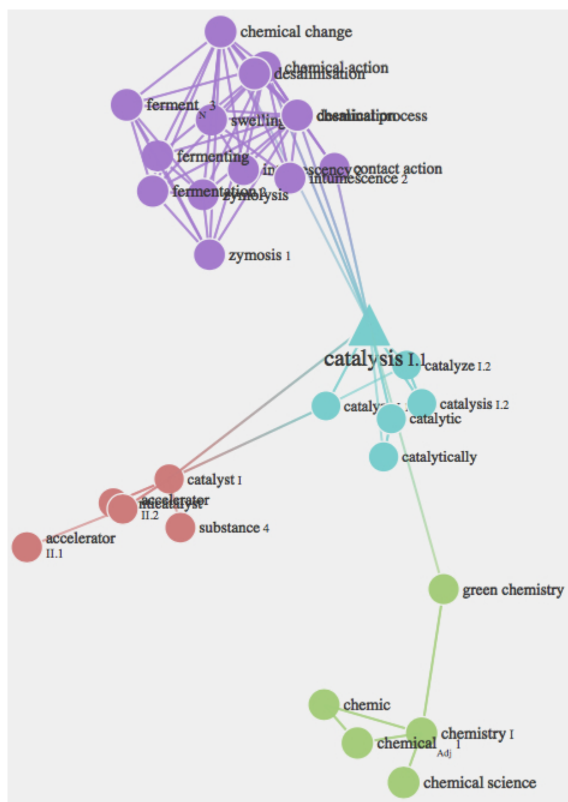


Figure 3: Semantic space of CATALYSIS **I.1**.

This semantic space features not only lexical units that are **directly** connected to CATALYSIS **I.1** – e.g. CONTACT ACTION or CATALYST **I** –, but also indirectly connected terms – e.g. GREEN CHEMISTRY or CHEMICAL CHANGE – that entertain significant semantic proximity with this headword.

At present, semantic space clustering is based

[7]The distinction between article-view and lexical graph perspectives on the en-LN bears some similarity with *written* vs. *graph information modes* in Pram Nielsen (2013).

of Proxemy analysis (Gaume, 2008). It is optimized by taking into consideration the *semantic weight* of each individual lexical function. For instance, a paradigmatic lexical function such as $\mathbf{S_1}$ possesses the maximal semantic weight "2" in the en-LN model of lexical functions, while the $\mathbf{Oper_1}$ lexical function denoting support verb collocates possesses the minimal semantic weight "0".[8]

We believe that lexical systems – small-world graphs of lexical units connected by paradigmatic and syntagmatic relations – are powerful alternatives to more traditional taxonomic models for at least two reasons: (i) they favor semantic space connectivity over a more restricted class-based organization and (ii) they unite both semantic and combinatorial connections within the same formal apparatus (lexical functions).

3.3 Definitional Embedding of Notions

To conclude, we wish to say a few words about definitions and their role in the structuring of terminological knowledge. As indicated in section 2, lexical definitions also participate in the weaving of lexical systems, though to a lesser extent, by implementing semantic embedding. In the specific case of CATALYSIS **I.1**, two lexical units appear in the article-view as clickable targets of definitional embedding links: REACTION **1** and the terminological idiom GIBBS ENERGY. This is made possible by the formal encoding the definition: an XML-like tagging of the definitional text, that we will not detail here for lack of space.

Acknowledgments

We are grateful to TIA 2015 anonymous reviewers for their comments on a preliminary version of our paper. This research is supported by the *PEPS Mirabelle 2015* program (CNRS and Université de Lorraine) for interdisciplinary research.

References

Yannick Chudy, Yann Desalle, Benoît Gaillard, Bruno Gaume, Pierre Magistry and Emmanuel Navarro. 2013. Tmuse: Lexical Network Exploration. In: *The Companion Volume of the Proceedings of IJCNLP 2013: System Demonstrations*, Asian Federation of NLP, Nagoya, 41–44.

Nabil Gader, Veronika Lux-Pogodalla and Alain Polguère. 2012. Hand-Crafting a Lexical Network With a Knowledge-Based Graph Editor. In: *Proceedings of the Third Workshop on Cognitive Aspects of the Lexicon (CogALex III)*, The COLING 2012 Organizing Committee, Mumbai, 109–125.

Nabil Gader, Sandrine Ollinger and Alain Polguère. 2014. One Lexicon, Two Structures: So What Gives? In Heili Orav, Christiane Fellbaum and Piek Vossen (eds.): *Proceedings of the Seventh Global Wordnet Conference (GWC2014)*, Tartu, 163–171.

Bruno Gaume. 2008. Mapping the Forms of Meaning in Small Worlds. *Journal of Intelligent Systems*, 23:848–862.

Roald Hoffmann and Pierre Lazlo. 1991. Representation in Chemistry. *Angewandte Chemie International Edition in English*, 30:1–16.

Marie-Claude L'Homme. 2002. Fonctions lexicales pour représenter les relations sémantiques entre termes. *Traitement automatique de la langue (T.A.L.)*, 43(1):19–41.

Veronika Lux-Pogodalla and Alain Polguère. 2011. Construction of a French Lexical Network: Methodological Issues. In: *Proceedings of the First International Workshop on Lexical Resources, WoLeR 2011. An ESSLLI 2011 Workshop*, Ljubljana, 54–61.

Alan D. McNaught and Andrew Wilkinson. 1997. *IUPAC. Compendium of Chemical Terminology (the "Gold Book")*, 2nd edition, Blackwell Scientific Publications, Oxford. On-line corrected version: `http://goldbook.iupac.org` (2006-) created by M. Nic, J. Jirat, B. Kosata; updates compiled by A. Jenkins.

Igor Mel'čuk. 1996. Lexical Functions: A Tool for the Description of Lexical Relations in the Lexicon. In Leo Wanner (ed.): *Lexical Functions in Lexicography and Natural Language Processing*, Language Companion Series 31, John Benjamins, Amsterdam/Philadelphia, 37–102.

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

Louise Pram Nielsen. 2013. User experimentation with terminological ontologies. In: *Proceedings of the 10th International Conference on Terminology and Artificial Intelligence TIA 2013*, Paris, 185–188.

Duncan J. Watts and Steven H. Strogatz. 1998. Collective dynamics of 'small-world' network. *Nature*, 393:440–442.

[8] There is no significant semantic link, in most cases, between a noun and its support verbs. The relationship between *flu* and its support verb *to have*, between *decision* and *to make* . . . is first and foremost combinatorial, not semantic.