# Exploring the Web of Coined Catchy Phrases

**Torbjörn Lager**
Department of Philosophy,
Linguistics and Theory of Science
University of Gothenburg, Sweden
torbjorn.lager@ling.gu.se

**Jenny Myrendal**
Department of Philosophy,
Linguistics and Theory of Science
University of Gothenburg, Sweden
jenny.myrendal@gu.se

## ABSTRACT

We employ a method inspired by corpus linguistics to illustrate the complexity and multifacetedness of the Web by extracting phrases about the Web from the Web itself. The phrases are divided into semantic categories and analysed with regards to their semantic content and rhetorical functions. We argue that the extracted phrases simultaneously have descriptive content and function as names. We also argue that they are used as rhetorical tools by actors from different groups in society involved in developing the Web, which is why we suggest that they need further attention by scholars interested in Web Science and linguistics.

## INTRODUCTION

The *World Wide Web* was the name given by Tim Berners-Lee in 1990 to what was later to become the Web as we now know it. Since then, lots of attempts have been made, by Berners-Lee and many others, to single out other particular aspects of the Web by coining phrases such as the *Semantic Web*, the *Web of People*, the *Semantic Web of Things* and many others. We believe there is something to be learned from such phrases since they reflect what people *have* thought, and still *think*, about the past, present and future Web. They are a phenomenon *on* the Web, they are *about* the Web, and they deserve to be studied.

Web Science is a science still looking for its methods. The method we are using in the present paper is inspired by methods used in the field of *corpus linguistics*, the study of language as expressed in samples of "real world" authentic text.

## COINED CATCHY PHRASES ON THE WEB

The phrases that we target match the patterns *the* <AP> *Web*, *the Web of* <$NP_{PL}$> and *the* <AP> *Web of* <$NP_{PL}$>, where <AP> is an adjective phrase and <$NP_{PL}$> is a noun phrase in the plural. Even outside titles and headings, words are often written with an initial capitalised letter (as indicated by the examples above), which seems to suggest that the phrases at least sometimes can be regarded as proper names. If not capitalised, they are sometimes "scare-quoted", indicating recognition of their use, but also a certain hesitance in using them.

To find authentic occurrences of such phrases we need access to very large corpora. There are indeed some large corpora available: The British National Corpus (BNC – 100 million words, 1980s-1993) and American National Corpus (ANC – 22 million words, 1990-) for example. Such corpora are balanced (i.e. samples are carefully selected in order to be representative), and often come with linguistically aware tools that allow their users to search for instances of patterns of the kinds that we are interested in. Unfortunately, these corpora are far too small as well as much too dated for our purpose.

In fact, the only possible source of text where a significant number of the kind of phrases that we are interested in are likely to exist is the Web itself. However, since it is not balanced, the Web is not a corpus in the technical sense of the word, nor is it part-of-speech tagged. Moreover, the tools for investigating the Web are definitely not up to the standards of state-of-the-art corpus linguistics tools.

For lack of a real corpus and adequate tools we choose to investigate the part of the Web that Google indexes, and we use Google Search as a corpus tool. Unfortunately, Google Search does not distinguish between upper and lower case letters, does not recognise quotation marks, and reported counts are not reliable (Kilgarriff, 2007). From a scientific point of view this is far from ideal and we tend to agree with Kilgarriff's conclusion that "Googleology is bad science". But in this case, bad science is better than no science, and we see no other way than to do our best with the tools Google provides. We have done so using search queries such as these:

```
"The * Web"
"The Web of *"
"The * Web of *"
```

By surrounding a phrase with double quotation marks we indicate that we are interested in the whole phrase rather than an unordered set of search terms. An asterisk matches one or more words.

Such queries allowed us to find a large number of the kind of coined and catchy phrases that we targeted, but only after a great deal of manual work. *Precision* was low since many phrases that match

the above patterns were deemed not to be of the right kind, and we had to filter them out by hand. Since we suspected (and were right to suspect) a low *recall* too, we generated, in a fairly systematic but still manual way, other candidates using the following simple heuristics: If an instance of "The Web of X" was already found, we looked for "The Web of Y" where Y belonged to the same "semantic field" as X. For example, since we had already found "The Open Web", we also searched for "The Closed Web", and since we had found "The Syntactic Web" and "The Semantic Web" we also searched for "The Pragmatic Web". We found such heuristics to be very fruitful.

We found no less than eighty-four phrases of the kind that we were looking for, presented in Table 1. For each phrase we recorded the counts as reported by Google but since low precision means that many of the results are irrelevant for our purpose, and since counts reported by Google are not reliable, we choose not to present them in the table. However, as an indication of the numbers involved, we note that the phrases are listed in falling order of frequency, with Mobile Web (61,400,000 results) being the most frequent and Haptic Web (2320 results) the least frequent. Again, we need to keep in mind that "the Mobile Web" is also matching the prefix of "the mobile web browser…" and that "the Signed Web" matches "the signed web advertising agreement", two examples of common kinds of constructions that explain the low precision. The important thing for us was to make sure that we had at least a handful of relevant matches, such as "the Mobile Web comes of age" and "to the reader of the Signed Web page, the hyperlink appears as a small video containing a sign or short phrase in Sign Language."

We also note that, with a few exceptions, relevant instances of the phrases also occurred in Google Scholar, the subset of the document indexed by Google that contains scholarly literature from all broad areas of research. Thus we can conclude that such phrases are not always just marketing hype.

| **Phrase 1-21** | **Phrase 22-42** | **Phrase 43-63** | **Phrase 64-84** |
|---|---|---|---|
| Mobile Web | Read/Write Web | Web of Documents | Web of Entities |
| World Wide Web | Social Semantic Web | Tagged Web | Conversational Web |
| Open Web | Semantic Web of Data | Physical Web | Semantic Social Web |
| Social Web | 3D Web | Two-Way Web | Transactional Web |
| Data Web | Collaborative Web | Virtual Reality Web | Hypertext Web |
| Deep Web | Read-Only Web | Written Web | Social Web of Things |
| Real-Time Web | Ubiquitous Web | Closed Web | Multimodal Web |
| Dynamic Web | Web of People | Centralized Web | Semantic Web of Things |
| Live Web | Web of Things | Spoken Web | Synaptic Web |
| Semantic Web | Hidden Web | 2D Web | Audible Web |
| Web of Trust | Multilingual Web | Spatial Web | Incremental Web |
| Personal Web | Surface Web | Cooperative Web | Decentralized Web |
| Desktop Web | Participatory Web | Indexed Web | Web of Events |
| One-Way Web | Invisible Web | Programmable Web | Syntactic Web |
| Visual Web | Static Web | Web of Services | Web of Sensors |
| Intelligent Web | Informational Web | Web of Linked Data | Web of Applications |
| Web of Data | Wisdom Web | Augmented Reality Web | Web of Places |
| Wireless Web | Pragmatic Web | Learning Web | Web of Devices |
| Geospatial Web | Visible Web | Emotional Web | Signed Web |
| Voice Web | Sensor Web | Semantic Sensor Web | Transient Web |
| Anti-Social Web | Classic Web | Contextual Web | Haptic Web |

**Table 1: Eighty-four coined and catchy phrases "describing" the Web. The underlined phrases indicate those occurring as a main header in a Wikipedia article (linked to in the online PDF).**

## THE WEB DESCRIBING ITSELF

Now that we have lots of them listed, what do such phrases tell us about (people's ideas about) the Web? The first thing to note is that they are very many and, as far as we know, many more than in other socio-technological domains. Perhaps Berners-Lee started a trend with his *World Wide Web*? Or maybe a web, understood as network or graph, just happens to be an abstraction applicable in very many domains?

Since they are so many, space does not permit us to deal with them all in the present paper, but it makes sense to give an overview. As regards their meaning, let us first note the obvious: the intended meaning of the same phrase may be different, depending on who is using it. Also, most of them are very vague and ambiguous. Furthermore, some phrases are used *synonymously*, thus indicating that the number of concepts is lower than the number of phrases. We have for example found no significant difference in meaning between the Cooperative Web and the Collaborative Web. Other phrases are *antonyms*, the Open Web and the Closed Web, for example.

With a syntactico-semantic generalisation we can say that the plural noun phrase in an expression of the form *The Web of* $<NP_{PL}>$ refers to the *kind* of entities that are linked, and that the adjective phrase in an expression of the form *The* <AP> *Web* ascribes a quality or property to the whole of the Web or to a part or an aspect of it. A phrase of the form *The* <AP> *Web of* $<NP_{PL}>$ does both. While this seems to suggest that our target phrases have descriptive contents, they are also, as noted above, sometimes used as names. However, nothing stops them from being *descriptive names*, i.e. referring expressions which have, unlike ordinary names, a descriptive content (Evans, 1982).

### Semantic categories and a Short Story of the Web

In this section, using (a sizable selection of) our coined phrases, we make an attempt to write a "story" of the Web, with different "chapters", seen as different semantic categories, illuminating different high-level aspects, such as linking, modality, communication, etc. It is not necessarily a *true* story, it is definitely not the *only* story, and the division into chapters may be done differently. Our intent is simply to demonstrate that our coined catchy phrases have a great deal of descriptive content.

**Linking**. The word *Web* is a metaphor (now dead) most likely deriving from the spider's web, reminding us that the Web is a *network* consisting of nodes and links connecting nodes with other nodes. In its most general form, it is a Web of entities such as documents, people, data, things, places and sensors.

Furthermore, people seem to be in some sort of agreement that the following near equivalences hold:

The Hypertext Web ≈ The Web of Documents
The Semantic Web ≈ The Web of Data
The Social Web ≈ The Web of People
The Physical Web ≈ The Web of Things
The Geospatial Web ≈ The Web of Places
The Sensor Web ≈ The Web of Sensors

Rather obviously, yet contrary to what the phrases seem to suggest, different phrases do not pick out *different* webs. Rather, they should be regarded as *aspects* or *facets* of the interconnected "mess" that is the one and only Web. Indeed, the *Web of Entities*, that seems to imply the interconnectedness of just about anything, might be a proper but not very informative name for the whole, of which the other webs form "sub-webs".

**Accessibility**. The *Indexed Web* refers to the portion of the Static Web that is indexed by (any of) the big common search engines such as Google or Bing. The *Surface Web* and the *Visible Web* appear to be other names for this part of the Web. The *Deep Web*, the *Invisible Web* and the *Hidden Web* are used to describe the portions of the Web that require login and password and/or the Dynamic Web of documents dynamically generated from databases, and which therefore cannot be indexed by "external" search engines capable of indexing the Static Web only.

**Modality**. As human beings we experience the world around us through *all* our senses, but the present Web is basically only capable of presenting itself in ways that affect our sight and hearing. The Classical Web was a Visual Web soon followed also by an Audible Web. The Haptic Web is not yet here (perhaps with the exceptions of online games capable of vibrating the players' smartphones or the pressure of a stylus on a graphics tablet connected to an online drawing application). The Visual Web of two-dimensional (2D) images is likely to evolve into a 3D Web of moving images, perhaps even into a Virtual Reality Web. The Web will then have to be a truly Multi-modal Web, presenting itself to all our senses and allowing human actuators (muscles) to manipulate the Web through haptic interfaces and machine sensors. The Audible Web of sounds, with output through loudspeakers and input via microphones is about to be refined into a Voice Web utilising speech recognisers and synthesisers, thus giving a whole new twist to the notion of a Conversational Web.

The Web is already equipped with sensors such as cameras and microphones, mirroring the human senses of sight and hearing. The Sensor Web will allow any physical phenomenon that can be detected and measured to provide input to the Web, including for example the radio waves emitted by RFID tags and GPS

satellites or the chemical compounds emitted by food about to go stale. On the basis of low-level quantitative sensor data the Semantic Sensor Web will be capable of providing both humans and machines with high-level qualitative (symbolic) descriptions of the status of the world around us.

**Ubiquity**. The Classic Web was a Wired Web and a Desktop Web, but the Mobile Web brought the Web to small and light devices such as smartphones. However, small size and light weight alone is not enough – for true mobility the Web must also be Wireless. If everybody is always carrying a device, always on and wirelessly connected to the Internet, the Web becomes Ubiquitous. Desktop computers and smartphones can already be regarded as forming a Web of Things (maybe best described as a Web of Devices), and once the Web of *other* Things becomes a reality, we will be constantly surrounded by things that are connected, and then the Web will become even *more* Ubiquitous.

**Communication**. The early Web was Read-only and One-Way in the sense that while anyone could in principle author and publish pages on the Open Web, only a few did, and the role of the producer (author) was usually clearly distinct from the role of consumer (reader). With the advent of wikis, blogs and social networking sites such as Facebook the Web became Read/Write and Two-Way and thus evolved into a Conversational Web. Another step towards a full-blown Conversational Web came with RSS and other technologies supporting a Real-Time Web allowing consumers to get notified in real-time when producers of their choice have produced something new. Finally, the integration of pre-web "conversational" technologies such as mail, chat and messaging into social networking sites has played a role too.

**Cooperation**. It is reasonable to regard communication as necessary for cooperation and therefore it makes sense to suggest that the Conversational Web provided the foundation for the Cooperative Web (and the Collaborative Web and the Participatory Web, which are here treated as synonyms to the Cooperative Web). Web "phenomena" such as Wikipedia and the open source software revolution would clearly not have been possible without communication.

**Intelligence**. Intelligence is a notoriously hard to define concept and there are a multitude of senses in which the Web can be said to be intelligent (or not). AI visionaries refer to a network of intelligent artificial agents capable of recommending the useful, extracting the essential, and automating the repetitive while others refer to the Web as a *whole* as intelligent, as a Synaptic Web, an emerging Global Brain or (more modestly) as a way to harness the collective intelligence of a web of humans and machines in symbiosis.

Intelligence entails reasoning and the Semantic Web supports mechanised reasoning so in that way the Intelligent and the Semantic Web are clearly related. Intelligence moreover entails sensitivity to context, and the Contextual Web is supposed to be a web that understands users and responds appropriately given the user's current context. Based on our situational context (where we are) and/or the recorded historical context of our previous decisions (how we used to behave) and/or our sentiments as we express them in our communications (what we say we like and dislike), recommendations and other kinds of information can be given that makes the Web appear as our Personal Web, a Web tailored to our interests, needs and wants. We only have to look at sites such as Amazon to get a taste of this.

**Politics**. Evolution does not have a built-in direction and the Web is not necessarily evolving the way we want. The Classic Web was also an Open Web where anyone could say anything about any topic. It was furthermore a Decentralized Web, with home pages being served from a multitude of servers owned by a multitude of individuals and very small to very large sized organisations. However, to ordinary users the Web was still Read-only since setting up a web server was technically too demanding for the majority of them. The Read/Write Web made it much easier for anyone to produce content, but usually only content to be stored on servers owned by someone else, content that can in principle be used for purposes beyond the producer's control. In contrast to the Open and Decentralized Web, the Closed Web is a Centralized Web of Applications where the owners of the applications, perhaps under the pressure of juridical law, perhaps guided only by their own systems of "cultural values and norms", can dictate what can be said and on what topic.

**Language**. Related to the Conversational Web, the Written Web, Spoken Web and Signed Web refer to uses of different forms of *language*. The Classic Web was Written and, by far, most of it still is, but the advent of the Mobile Web with devices too tiny for user-friendly keyboards, with the availability of standards such as VoiceXML, with other speech related standards in the works, and with applications such as Apple's Siri, the Spoken Web (aka the Voice Web) seems to be gaining traction. Meanwhile, video in combination with "sign-linking" web technology will perhaps spark a future incarnation of the Web that is driven entirely by sign language content, thus giving the deaf community its own Signed Web.

The notion of a Multilingual Web seems to reflect both a fact and a hope: the fact that web content has always been authored in different languages, and the hope that with the advent of machine translation technologies this will become less of a *problem* in the future.

### Preliminary discussion

Our exercise above sorted the phrases into a fairly small set of categories into which they seem to cluster fairly naturally. The categories are listed in Table 2.

| | | |
|---|---|---|
| **Linking** | **Ubiquity** | **Intelligence** |
| **Accessibility** | **Communication** | **Politics** |
| **Modality** | **Cooperation** | **Language** |

**Table 2: The nine semantic categories into which we have sorted most of the catchy coined phrases**

This is of course very tentative since we can easily imagine other ways to cluster the phrases. One might want to merge the categories of Communication and Language for example. Still, we are impressed by the way sensible high-level facets of the Web emerged from lower ones, backed up by findings in corpus data.

At one point we nearly fell for the temptation to introduce a Linguistics category, comprising the Syntactic Web, the Semantic Web and the Pragmatic Web. However, we decided that we prefer to regard linguistics not as a facet of the Web, but as an approach to the study of the facets Language and Communication.

## THE WEB PERSUADING ITSELF

The language choices we make are the foundation of our communication with other people and reflect how we experience and interpret different aspects of the world around us. So what roles might the phrases presented in this paper play in the discourse of web technology? Given that the introduction of many other technological innovations historically has been preceded by and immersed in influential rhetorical discourse (Barry, 1991; Miller, 1994; Coyne, 1995; Johansson, 1997; Almqvist, 1998), it is reasonable to believe that the different expressions that reflect the complexity of the Web also serve some kind of rhetorical function in the discourse where they are being used.

In Technobabble from 1991, Barry was among the first to describe a relationship between the evolution of computer technology and its associated lexicon. Barry defines technobabble not just as meaningless chatter about technology, but as an important communicative factor that influences actors working in development of the rapidly growing technology industries. Miller, in a well-noticed article in Argumentation from 1994, analyses rhetorical devices found in technological forecasting literature and arrives at the conclusion that how we talk about technology operates as "technological forecasting" in the way that it defines the future of how technology and society is shaped. Miller defines technological forecasting as "a discourse in which the characterisation and construction of moments in the present are crucial to the projection of the future" (Miller, 1994, p 82). Miller means that by analysing how different actors "sell" their visions of the future, we can learn a lot about how society is shaped and how technology develops as part of a society. Bazerman (1998) also highlights the important role of rhetoric in technological development, in describing a dialectic relationship between rhetoric and technology. He suggests that the "rhetoric of technology shows how the objects of the built environment become a part of our systems of goals, values and meaning, part of our articulated interests, struggles and activities". (Bazerman, 1998, p 386).

Along the same lines, in a dissertation from 1997, Johansson was able to show how the language use that surrounded the introduction of new computer technology between 1995 and 1995 in Sweden primarily served a rhetorical function, as different actors (producers, users, critics/propagators and politicians), put forth different arguments (political, social, technical and economic) to speed up the introduction of new infrastructures for technology. Johansson concludes that the words and phrases used to characterise new technology creates images which later influences how new products are designed, used and perceived: "Contemporary visions and beliefs are projected onto technology, and influence not only how technology in itself is conceived and/or interpreted, but also what kind of technology is believed to be useful for "solving" problems, both close at hand and for a society as a whole. Different actors propagate different solutions, and try to "win" others for their cause, primarily by the use of rhetorical devices in the form of "texts" of all kinds. At the same time, technology helps to set the frames of our minds, thus being formative with regard to how we think and feel about ourselves and society. In this way, technology serves as a "mindsetter" for our present society and for visions of the future." (Johansson, 1997, p 213).

Following Johansson, Almqvist (2001, p 14) also stresses that how we choose to talk about the Internet gives rise to discursive meanings about how technology functions (or should function) in society.

To attract attention to one's own arguments from listeners or readers in a communicative setting, an actor can employ various rhetorical devices, which Johansson defines as "linguistic tools and kinds of arguments that are used to make the argumentation efficient" (Johansson, 1997, p 52). For example, talking about aspects of the Web as being Open or Closed

is part of a political rhetorical discourse, whereas drawing attention to the aspects of the Web as either Syntactic or Semantic can be seen as part of a scientific or technological rhetoric. To coin and introduce two concepts simultaneously seems to be a commonly used rhetorical approach, when the objective is to draw attention to a certain aspect of the Web. For example, the phrases Read-Only web and Read/Write web were introduced around the same time to highlight the interactive and dynamic aspects of the Web. Nobody ever talked about the Web as being Read-Only until the Web became more interactive, so coining the two phrases at the same time can be seen as a rhetorical attempt to create a contrasting backdrop, against which the new concept (Read/Write) is positioned.

Naturally, it is impossible for us to perform an exhaustive rhetorical analysis based solely on the phrases above taken out of context. However, we believe that it is likely that these expressions are used (at least partially) with a rhetorical purpose by different societal groups, to draw attention to various aspects of web technology that these groups perceive as the most important for future technological development. Therefore, it becomes important to study the interchange between these actors and their different ways of putting forward their own technological images and agendas by analysing language choices, and how they in turn influence the technological evolution. Accordingly, the phrases described in this paper might be said to function as a forward-striving rhetorical force in current and future web development, which is why those of us interested in Web Science and linguistics should pay them closer attention.

## SUMMARY AND CONCLUSIONS

The Web is an exceedingly fast-evolving, complex and multifaceted technological artefact and social phenomenon. This alone may explain why the coined and catchy phrases are so many, and why their roles vary.

They are *names*, singling out and referring to different parts or aspects/facets of the Web. Regarded as referring expression, we note that they may initially not refer to anything existing in the real word, but rather to express a *vision* of a future Web, an idea of something to be created. The phrases seem to have a "sticky" quality, since once something is created they tend to be used regardless of how well they continue to express the initial vision. As names, they are often used in combination with other words, as in "Semantic Web vision", "Sensor Web community" and "Open Web movement".

They have *descriptive content* – describing the Web from different angles. We believe that we have shown that by placing them in a narrative context of phrases in the same and related "semantic fields" one is able to create a common-sense "story" of the Web, with "chapters" illuminating some of its higher-level aspects, such as linking, modality, etc. Such a story may not be entirely true, and it greatly simplifies things, but it seems to capture people's main ideas about the Web.

They are *rhetorical tools* in the hands of scientists, technologist, marketers and politicians to persuade research-funding agencies, technology investors and ordinary users of the Web to invest energy, time and money into the development or this or that aspect of the Web. They are often introduced in pairs, where the strength of a new concept is illustrated by contrasting with the (now) perceived weakness of the "old" Web. The Web is a Web of Rhetoric, a web talking to itself, about itself, trying to persuade itself to move in this or that direction.

Web Science must not be caught in the Web of Coined and Catchy phrases. Like any science, Web Science should, as much as possible, wrap itself in a mantle of disinterested curiosity. But looking at it, and describing it, is important, since it influences the way the Web evolves.

Finally, we will definitely claim a place for linguistics and rhetoric in the next version of the Web Science "butterfly".

## ACKNOWLEDGMENTS

This paper has been written within the project *The Web: A view from (computational) linguistics,* funded by the Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg.

## REFERENCES

Almqvist, J. (2001). Bilder av Internet – en studie av IT som verktyg för meningsskapande. *Utbildning och demokrati, 10*(1), 7-27.

Barry, J. A. (1991). *Technobabble*. Cambridge, Mass: MIT Press

Bazerman, C. (1998). The Production of Technology and Production of Human Meaning. *Journal of Business and Technical Communication, 12*(3), 381-387.

Coyne, R. (1995). *Designing Information Technology in the Postmodern Age. From Method to Metaphor.* Cambridge, Mass: MIT Press

Evans, G. (1982). *The varieties of reference*. Oxford: Oxford University Press.

Johansson, Magnus (1997): *Smart, Fast and Beautiful. On Rhetoric of Technology and Computing Discourse in Sweden 1955-1995.* Linköping Studies in Arts and Science 164. Linköping University.

Kilgarriff, A. Googleology is bad science. *Computational Linguistics* 33 (1): 147-151, (2007).

Miller, C. R. (1994). Opportunity, opportunism, and progress: Kairos in the rhetoric of technology. *Argumentation, 8*(1), 81-96.