Finding gaps in semantic descriptions. Visualisation of the cross-reference network in a Swedish monolingual dictionary

Kristian Blensenius¹, Emma Sköldberg², Erik Bäckerud³

¹ University of Gothenburg, Gothenburg (Sweden)

² University of Gothenburg, Gothenburg (Sweden)

³ The Swedish Academy Dictionary, Lund (Sweden)

E-mail: kristian.blensenius @gu.se, emma.skoldberg@svenska.gu.se,

erik. backerud@svenskaakademien.se

Abstract

Providing lexical information in dictionary entries by cross-referencing between semantically related headwords is very important, both from a reception-oriented and a production-oriented perspective. This study presents a survey of cross-references in a comprehensive monolingual dictionary of Swedish. It discusses cross-referencing in dictionaries in general as well as in the Swedish dictionary, focusing on the following four types of paradigmatic cross-references: SEE, COMPARE, SYNONYM, and OPPOSITE. By using data-visualisation software, the semantic network in the dictionary is overviewed in a new way. Furthermore, errors, gaps as well as other areas of improvement in the dictionary related to cross-referencing are discovered. Moreover, the relationships between the existing cross-references, how they are introduced in the dictionary and the dictionary's intended target groups are addressed. The study also reveals that the traditional lexicographic policies of the dictionary need to be adjusted to take advantage of the transition from paper to electronic publication.

Keywords: cross-references; paradigmatic relations; Swedish; lexicography; semantics

1. Introduction

According to Atkins and Rundell (2008: 238), "every dictionary has its own palette of admissible ways of cross-referring from one entry to another". The main aim of this study is to present a survey of cross-references in a comprehensive monolingual dictionary of modern Swedish, in this case the second edition of *Svensk ordbok utgiven av Svenska Akademien* ('Contemporary Dictionary of the Swedish Academy'; henceforth SO). The study is conducted using a software for visualisation of graph data. We will show how the software can be used to find errors, gaps as well as other areas of improvement in the dictionary.

The outline is as follows: In section 2, the different functions of cross-references in general are discussed. Section 3 presents the different types of cross-references in SO. Section 4 focuses on the usage of four different paradigmatic cross-references placed in special fields in the microstructure of the dictionary and on how they can be visualised. These cross-references are indicated by the labels SE ('see', henceforth: SEE), JFR (abbr. of *jämför* 'compare', henceforth: COMPARE), SYN. (abbr. of *synonym* 'synonym', henceforth: SYNONYM), and MOTSATS ('opposite', henceforth: OPPOSITE). Section 5,

finally, gives some final remarks.

2. Cross-references in dictionaries

Information on semantically related headwords in the dictionary entry is very important, especially from a reception-oriented perspective. Synonyms, antonyms, etc. serve to provide access to additional lexical information and lexical sets (e.g. relations between nouns such as *north*, *south*, *east*, and *west*) as well as delimiting the meaning of the headword (see e.g. Järborg, 1989: 20; Hult et al., 2010). For instance, the meaning of food-related English verbs like *chop*, *grind*, *mash*, and *shred* becomes clearer when comparing the definitions of the words. This kind of information can also serve to enhance users' knowledge of connotations, pragmatic characteristics, etc. For example, users can, by comparing different entries, be made aware of the different emotive meanings of adjectives denoting (degrees of) overweight, e.g. *chubby*, *corpulent*, *fat*, *plump*, and *stout*.

In her classical work Words in the Mind. An Introduction to the Mental Lexicon, Aitchison (2003) discusses four types of relationships between stimulus words and response words in association tests; co-ordination (e.g. salt - pepper), collocation (e.g. bright - red), superordination (e.g. color - red/blue/green), and synonymy (e.g. hungry - starved) (ibid. 2003: 84–91). She states that the consistent answers given in association tests seem to testify that meaning relations between different words also have psychological validity. Related words seem to be stored so that they form a system within which the associations take place.

Furthermore, information about semantically related headwords is, according to Malmgren (2009: 98), extremely important from a *production*-oriented perspective; providing synonyms can help users to write or speak with a varied vocabulary. Moreover, knowledge of antonyms and other classes of converse pairs of words can also be useful when it comes to paraphrasing (cf. *not nervous* and *calm*).

3. SO

3.1 The second edition of SO and its precursors

The second edition of SO, published in spring 2021, has been compiled at the Department of Swedish at the University of Gothenburg. The dictionary is primarily based on three previous (printed) dictionaries, including the first edition of SO, released in 2009. The dictionary includes approx. 65,000 headwords. In short, the monolingual definition dictionary is descriptive, and it deals with contemporary general language. SO is mainly intended as a reception dictionary, but it can also be used for production, and the target user groups are native speakers and advanced L2 learners.

The second edition of SO will (in contrast to the first edition from 2009) only be published digitally, as a dictionary app and at the Swedish Academy dictionary portal Svenska.se. In preparing the second edition, the editorial team has made efforts in using the digital format as much as possible. For example, the content of the dictionary has become more accessible to the users than before, by having been made searchable in different ways. Furthermore, the SO lexicographers have tried to make it easier for users to both review the contents of the entries and go from one entry to another by adding more cross-references (these are indicated by labels, e.g. SYNONYM) and links.

At the same time, the current lexicographic team has updated the editorial guidelines with regards to cross-references between different entries. On many occasions, e.g. in the case of the verb 'die', there are plenty of synonyms or near-synonyms. However, when it comes to Swedish, there is a good selection of synonym dictionaries (see e.g. the website Synonymer.se, which is probably the most used Swedish dictionary of today). For this reason, and due to time constraints, inclusion of synonyms in SO has not been prioritised in the revision of the dictionary.

Furthermore, in connection with the practical lexicographic work with so-called "controversial" words, the SO lexicographers have aimed to include cross-references from offensive headwords to more neutral headwords in the dictionary, but not the other way around. For instance, in the second edition of SO, there is a link from the slightly archaic and derogatory adjective **homofil** ('homo') to the more neutral adjective **homosexuell** ('homosexual'). However, there is no cross-reference from the headword **homosexual** to **homofil** (see Petersson & Sköldberg 2020 for more details on this work; also cf. "reciprocal cross-references" and "one-way cross-references" in Svensén 2009:389).¹

3.2 Different types of microstructural cross-references in SO

In the following, we will discuss different types of cross-references included in the microstructure of the second edition of SO.

As a starting point, let us look at the entries ${}^{1}\mathbf{aktiv}$ ('active') and ${}^{2}\mathbf{aktiv}$ ('active voice') in the web version of SO in Figure 1.

¹ We use bold to indicate headwords. All translations into English were made by the authors.

publice	rad: 2021
¹ aktiv aktivt aktiva	
ORDKLASS: adjektiv	
UTTAL: ak'tiv 📣	
1 som deltar på ett verkningsfullt sätt i allmänhet el. i viss verksamhet	
MOTSATS inaktiv, ¹ passiv 1 JFR livaktig, verksam 1	
	DÖLJ –
SAMMANSÄTTN/AVLEDN.: nattaktiv	
KONSTRUKTION:	
aktiv (i nÅgot)	
EXEMPEL: hon var mycket aktiv i miljödebatten; klubben har 100 aktiva och 50 pas medlemmar	siva
○ särskilt som är verksam i yrkesliv eller idrott	
EXEMPEL: lagets målspottare beräknar att vara aktiv ett par år till	
○ äv. om handling och dylikt	
EXEMPEL: ett aktivt liv; aktivt föräldraskap	
○ äv. om naturföreteelser, bl.a. radioaktiva ämnen som utvecklar energi	
SAMMANSÄTTN/AVLEDN.: radioaktiv	
EXEMPEL: seismiskt aktiv; aktiva komponenter	
aktiv dödshjälp	
SE dödshjälp	
aktivt kol	
SE kol	
aktivt ordförråd	
SE ordförråd	
HISTORIK: belagt sedan 1839; av lat. <i>activus</i> 'praktiskt verksam', till <i>ag'ere</i> 'handla ursprung till agera	l; jfr
2 som formellt anger subjektet som i någon mening handlande om verb(form)	
MOTSATS ¹ passiv 2	
	DÖLJ –
EXEMPEL: i satsen "katten jagar råttan" har verbet "jagar" aktiv form	
○ āv. om motsvarande sats	
HISTORIK: belagt sedan 1820	
² aktiv aktivet aktiver	
ORDKLASS: substantiv	
UTTAL: aktiv 📣	
en verbform som normalt markerar att satsens subjekt är upphov till verbhandlingen	
MOTSATS ² passiv	DÖLJ –
HISTORIK: belagt sedan 1801; se ursprung till ¹ aktiv	

Figure 1: The entries ¹aktiv and ²aktiv in the web version of the second edition of SO.

Starting from the top of the entry ¹**aktiv**, there are four different paradigmatic crossreferences. The headwords **inaktiv** ('inactive') and ¹**passiv 1** ('passive') are introduced by the OPPOSITE label (i.e. MOTSATS in the figure). Furthermore, the users are encouraged to COMPARE the headword (JFR in the figure) with the adjective headwords **livaktig** ('lively') and **verksam 1** ('energetic'). It should be noted that in the dictionary two other types of paradigmatically related words are also included, namely the ones introduced by the labels SEE and SYNONYM (see section 4.2 below).

Second, there are cross-references in the idiomatic-expressions section (and other kind of multiword expressions). During the preparation of the first edition of SO, new guidelines were set up for the lemmatisation of idioms. For instance, an idiom including an adjective and a noun (e.g. *den röda tråden* 'the common thread, theme') was placed and defined in the noun entry with a cross-reference to the expression in the adjective entry. In the case of **1aktiv**, we find three cross-references of this type (see e.g. the multiword expression *aktiv dödshjälp* 'active euthanasia' with cross-reference to the entry **dödshjälp**).

As already indicated, the first edition of SO was primarily a printed dictionary, and the reason for using cross-references between idioms was limited space (cf. Rundell 2015). However, as the second edition of SO is only published electronically, it is at least possible to present the same information on an idiom in more than one entry, instead of using cross-references. A more radical change would be to make the idioms in SO more independent/visible/searchable and less dependent on their constituent parts. In this case, the traditional lexicographic policies of SO need to be adjusted to take advantage of the change in publication format.

Finally, as indicated by Figure 1, there are also links among the etymologies. In the case of ${}^{1}\mathbf{aktiv}$, the dictionary users are (more or less) informed that they should compare the historical information of this word with the etymology of the verb **agera** ('act').

In ²**aktiv** ('active voice'), there is one cross-reference and one link, to the OPPOSITE ²**passiv** (also a verb form) and to the etymology of the adjective ¹**aktiv**, respectively.

In addition to the cross-references and links found in the example in Figure 1, some words in the SO definitions are hyperlinked. One example is the noun entry **feminist** with the definition 'followers of feminism' where the word *feminism* is hyperlinked. These links are new in relation to the first edition from 2009, and they form also an important part of the semantic network in the dictionary. At present there are 13,000 links of this type in SO, so they are still relatively few. This can be related to the situation in English dictionaries. For example, Rundell stated already in 2015 that "In the digital editions of most (if not all) of the British learner's dictionaries, every word in an entry is hyperlinked" (Rundell, 2015:315). According to the author, it is advantageous that users can rapidly find the entry for a word if they are unsure about any word in a definition. However, according to the same author, it is never ideal for the users to have to go from one entry to another in order to get the full picture. If the meaning of certain keywords in the SO definitions (such as *feminism* in the example above) is made explicit in more than one entry (e.g. in the adjective entries **feminist** and **feministisk** 'feminist'), this problem can possibly be reduced.

To sum up, there are cross-references and links in slightly different places in the entries

¹**aktiv** and ²**aktiv**, and they partly serve different functions. Some are intended to facilitate for users to grasp the meaning of the headword by relating them to other relevant headwords in the dictionary. Other cross-references are only "space savers". The question is whether users understand this, if the labels are sufficiently informative, etc. (also see section 4.2 below).

4. The cross-reference network in SO

The tool we use to produce the graphics in this text is called Constellation (https://www.constellation-app.com/). It is a free open-source software for data visualisation and analytics originally developed at the Australian Signals Directorate in 2012.

To obtain the data describing all cross-references as a graph, the presentation-ready HTML files for the dictionary have been scanned by a Java program that generates a comma-separated text file with one line for every reference. This is one of the input formats accepted by the Constellation program. Each line contains identifiers and part of speech for the source and destination headwords as well as information about the kind of reference being made. An advantage of this simple data representation is that it makes it easy to select different subsets of the data for use in specific cases.

4.1 Visualisation of the cross-reference network in SO

The second edition of SO includes approx. 76,000 cross-references and links of the types discussed in section 3.2. The visualisation format is very space consuming, and some entries are connected to other entries by a great number of cross-references. For instance, **gå** ('go', verb) includes no less than 88 cross-references and links. Other examples from SO are **hand** ('hand', noun, 75 cross-references and links), **dag** ('day', noun, 71), **fot** ('foot', noun, 55), **stå** ('stand' verb, 54), and **dra** ('pull', verb, 54). The large number of cross-references and links in these entries is related to the fact that the words occur in a significant number of idioms.

Some examples from SO to illustrate how the software works and can be used in practical lexicography are presented below. The relationship between the entries ¹**aktiv** and ²**aktiv** and (a subset of) other headwords in the second edition of SO can be visualised as in Figure 2.

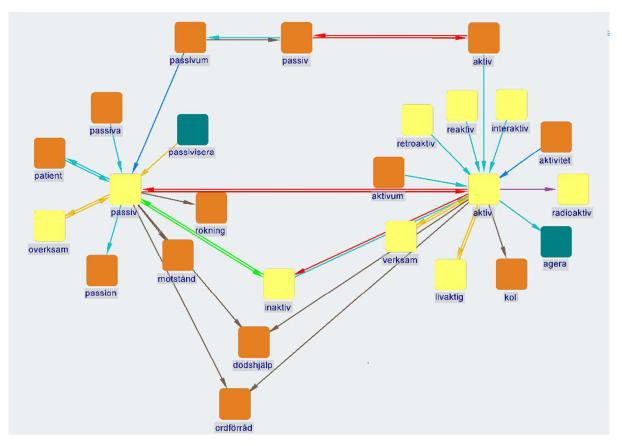


Figure 2: The semantic network including the headwords ¹**aktiv** and ²**aktiv** in the second edition of SO, illustrated by the Constellation software.

It can be seen that the figure includes boxes with different colours. Yellow boxes represent adjectives, orange boxes represent nouns, and green boxes represent verbs. The direction of the arrows indicates the direction of the cross-references. Furthermore, the colours of the arrows indicate the type of cross-reference. A green arrow represents SYNONYM, a red arrow represents OPPOSITE and a yellow represents COMPARE. Brown arrows represent SEE, a label used both in the section including paradigmatic cross-references and in connection with the idioms. In addition, light blue arrows represent hyperlinked words in the etymological part of the entry and dark blue arrows represent hyperlinked words in definitions (see e.g. the example entry **feminist** in section 3.2).

From the lexicographer's perspective, Figure 2 shows that many entries are treated in a consistent way. For example, there are cross-references in both directions (i.e. reciprocal cross-references) between the adjective headwords **aktiv** and **passiv** and between the noun headwords **aktiv** and **passiv**. As indicated by the colours of the arrows, the cross-references between the word pairs are also of the same type. However, the figure also reveals some shortcomings in the dictionary. For instance, there is a cross-reference (OPPOSITE) from **aktiv** to **inaktiv**, but it is unidirectional.

Finally, through the visualisation it becomes clear that a large proportion of the entries

in SO includes no cross-references at all. Most of these entries could easily be enhanced in different ways. For example, the entry **alarmerande** ('alarming') could provide information about more or less synonymous adjectives (e.g. **skrämmande** 'scary'). Furthermore, an entry such as **överviktig** ('overweight') could be enhanced by providing an antonym such as **underviktig** ('underweight'). Likewise, in the entry **animalisk** ('animal-like'), a cross-reference (COMPARE) to the cohyponym **vegetabilisk** ('vegetable-like') could be included. In addition, the content of many entries may become more accessible if individual words in the definitions are hyperlinked. In other words, by using illustrations like the one in Figure 2, lexicographers find different kind of gaps in the lexical database.

4.2 Paradigmatic cross-references in SO

In this section, we will focus on the usage of paradigmatic cross-references placed in certain fields in the microstructure of the dictionary, i.e. the cross-references indicated by the (translated) labels COMPARE, SYNONYM, SEE, and OPPOSITE (also cf. the types mentioned by Svensén 2009:248–251). In total, there are about 38,000 cross-references of this type in the dictionary. The total number is distributed as follows on the four current types: 26,011 COMPARE, 3,746 SYNONYM, 3,552 SEE, and 1,286 OPPOSITE. As already mentioned, the relationship between the headwords connected by cross-reference is indicated by the colour of the arrow. The yellow arrow represents COMPARE, and the green arrow represents SYNONYM. Furthermore, the brown arrow represents SEE and the red arrow represents OPPOSITE.

Before discussing more specific headwords, the choice of labels in SO should be considered. First, the use of abbreviations (i.e., JFR and SYN.; see section 1) is disputable. In order to make things easier for users, especially for L2 learners, these abbreviations should be expanded in future editions of SO. However, even if the labels are expanded, it is not obvious how labels such as COMPARE, SYNONYM, SEE and OPPOSITE should be interpreted. If the dictionary users follow cross-references marked with COMPARE and SEE, they might notice the semantic relationship between the words. Perhaps they also understand the intention of connecting to these closely related words (see section 2 above). However, the SO lexicographers have limited knowledge about this, and a user survey is needed to gain in-depth knowledge of this aspect.

Furthermore, the term *synonymy* usually denotes a meaning relation in which a word can be said to have the same meaning as another word. Principally, they should then have the same set of semantic components. This does not necessarily mean that the words are fully interchangeable in all contexts. Or, as Ullmann stated already in 1962 (p. 141), "In contemporary linguistics it has become almost axiomatic that complete synonymy does not exist". However, the term is common in teaching contexts, especially in second language teaching where it often refers to words that have approximately the same meaning. As already mentioned, there are more than 3,700 SYNONYM cross-references in SO, and a quick look at the words shows that they do not mean exactly the same thing. The two headwords categorised as synonyms often belong to either general or technical language and they often appear in different contexts. In the same way, the synonyms may differ in terms of style. It should also be mentioned that the SO editorial staff, in compiling the second edition, has been very restrictive regarding the use the label SYNONYM, since the headwords are seldom completely interchangeable. Instead, the vaguer COMPARE has been used. In this case, it is possible to discern a change over time concerning the principles for the lexicographical work with SO and its forerunners (and possibly also of the boundaries of the meaning of words like *synonym*). It is not satisfactory that there are differences between the entries in SO depending on the period the entries were compiled.

Thanks to the visualisation software, it is easier than previously to get an overview of the semantic network in SO. Inconsistencies and opportunities for improvement in the dictionary are also more obvious. As an example, the headwords related to the adverb **ganska** ('quite, pretty, rather') are presented in Figure 3.

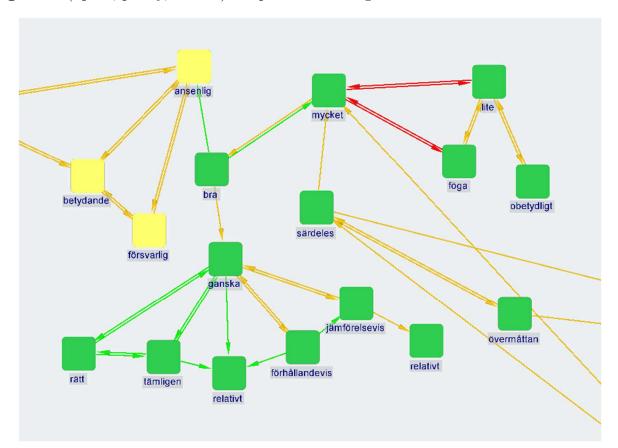


Figure 3: The headword **ganska** and its synonyms etc. in the second edition of SO.

As shown in Figure 3, there are reciprocal cross-references between **ganska** and two other headwords in SO (see the double green synonym arrows to and from the words **rätt** 'fairly' and **tämligen** 'fairly, moderately'). Furthermore, there are reciprocal yellow cross-references of the type COMPARE between **ganska** and **förhållandevis** ('proportionately') and between **ganska** and **jämförelsevis** ('comparatively'). The adverb **relativt** ('relatively') is also introduced as a synonym to the actual headword. The alternatives presented to the current adverb are, as already mentioned, important from a production-oriented perspective (see section 2).

Moreover, according to Figure 3, there is no cross-reference from the headword **relativt** to **ganska** – or from **relativt** to any other headword in the dictionary. In other words, the adverb **relativt** could be more clearly related to other adverbs in SO.

In the same figure, it is also possible to overview the cross-references to and from the headword **mycket** ('very'). According to the dictionary, one of the main senses of the adverb **bra** ('to a large extent') corresponds to a sense of **mycket**. The adverb **mycket** is, however, not considered as a synonym to **bra**. The relationship between **bra** and **mycket** is not clarified, but users are encouraged to compare the words. The information provided in the two entries is thus not entirely consistent.

Finally, as shown by Lyons (1977:270–290) among others, it is complicated to discuss the label OPPOSITE because it includes rather disparate categories. The most important types of opposite relations are probably *complementarity* (e.g. the words *dead - alive*), *antonymy* (e.g. *hot - cold*) and *converseness* (e.g. *husband - wife*). As with synonyms, the label OPPOSITE is used in a broad sense in SO, covering at least complementarity and antonymy. And, like synonyms, there are plenty of cases where the information on presumed converse headwords is not consistent.

5. Final remarks

In this paper, we focus on cross-referencing in the second edition of the monolingual contemporary dictionary *Svensk ordbok utgiven av Svenska Akademien* (SO). The cross-references, which create a semantic network, have been investigated by using a software for the visualisation of graph data.

As already mentioned, in preparing the second edition the editorial team has made efforts to use the digital format as much as possible. However, significant work remains before the digital format is fully utilised in SO, especially when it comes to the use of cross-references in the dictionary. A large proportion of entries in SO are completely isolated as they include no cross-references or links to other entries at all.

Overall, there is room for improvement concerning the semantic relations in the dictionary, which are unveiled by the visual network. Hopefully, a more developed version of the visualisation tool could be incorporated into the lexicographers' editing interface.

Furthermore, the relationship between the paradigmatic cross-references and the dictionary's intended target groups and intended areas of use can be discussed. Mother-tongue speakers and advanced learners do not always have the same possibilities for interpreting the information given. In addition, semantically related words that are

supposed to support reception and production, respectively, are not necessarily the same words. Moreover, the idea of, for example, 'synonymy' has probably changed over the approx. 40 years that the work with the dictionary database has been going on. The fact that the SO, from now on, is only published electronically can also significantly affect the information category in future editions.

In the future, it would be interesting to compare the semantic network in SO with the networks in other dictionaries. In the same way, it would be of interest to compare the network with other presentations of semantically related word. For Swedish, Åke Viberg and his colleagues have compiled a WordNet, but unfortunately it is not very accessible (see Viberg et al., 2002). A more updated resource is Swesaurus, which has been developed at Språkbanken at the University of Gothenburg (see e.g. Borin & Forsberg, 2014).

Finally, it would also be interesting to investigate whether the user interface could be provided with illustrations to further clarify how the Swedish words are related to each other, from a contemporary as well as from a historical perspective.

6. References

- Aitchison, J. (2003). Words in the Mind. An Introduction to the Mental Lexicon. 3rd edition. Oxford: Basil Blackwell.
- Atkins, S.B.T. & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.
- Borin, L. & Forsberg, M. (2014). Swesaurus; or, The Frankenstein Approach to Wordnet Construction. In: Proceedings of the 7th Global WordNet Conference (GWC 2014). Tartu: Global WordNet Association, pp. 215–223.
- Hult, A.-K., Malmgren, S.-G. & Sköldberg, E. (2010). Lexin a report from a recycling lexicographic project in the North. In A. Dykstra & T. Schoonheim (eds.) *Proceedings of the XIV Euralex International Congress*. Leeuwarden/Ljouwert: Fryske Academy, pp. 800–809.
- Järborg, J. (1989). Betydelseanalys och betydelsebeskrivning i Lexikalisk databas. (Unpublished manuscript.) Göteborg.
- Lyons, J. (1977). Semantics. Vol. 1. Cambridge: Cambridge University Press.
- Malmgren, S.-G. (2009). On production-oriented information in Swedish monolingual defining dictionaries. In S. Nielsen & S. Tarp (eds.) Lexicography in the 21st Century. In honour of Henning Bergenholtz. (Terminology and Lexicography Research and Practice 12.) Amsterdam/Philadelphia: John Benjamins, pp. 93– 102.
- Petersson, S. & Sköldberg, E. (2020). To discriminate between discrimination and inclusion: a lexicographer's dilemma. In Gavriilidou, Z. et al. (eds.) Proceedings of XIX EURALEX Congress: Lexicography for Inclusion, Vol. I. Alexandroupolis, Greece: Democritus University of Thrace, pp. 381–386.
- Rundell, M. (2015). From Print to Digital: Implications for Dictionary Policy and

Lexicographic Conventions. Lexikos 25, pp. 301–322.

- SO = Svensk ordbok utgiven av Svenska Akademien (First ed. 2009, second ed. 2021). Stockholm: Norstedts. Also available as apps and at https://svenska.se/.
- Svensén, B. (2009). A Handbook of Lexicography: The Theory and Practice of Dictionary-Making. Cambridge: Cambridge University Press.
- Ullmann, S. (1962). Semantics. An Introduction to the Science of Meaning. Oxford: Basil Blackwell.
- Viberg, Å, Lindmark, K., Lindvall, A. & Mellenius, I. (2002). The Swedish WordNet Project. In Braasch, A. & Povlsen, C. (eds.) Proceedings of the X EURALEX Congress: Copenhagen-Denmark, August 13-17, 2002. Proceedings Volume I, pp. 407–412.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

http://creativecommons.org/licenses/by-sa/4.0/

