https://doi.org/10.14618/ids-pub-11193

Nico Dorn

AN AUTOMATED CLUSTER CONSTRUCTOR FOR A NARRATED DICTIONARY

The Cross-reference Clusters of *Wortgeschichte digital*

Abstract *Wortgeschichte digital* (Digital Word History) is an emerging historical dictionary of the German language that focuses on describing semantic shifts from about 1600 through today. This article provides deeper insight into the dictionary's "cross-reference clusters," one of its software tools that performs visualization of its reference network. Hence, the clusters are a part of the project's macrostructure. They serve as both a means for users to find entries of interest and a tool to elucidate relations among dictionary entries. Rather than delve into technical aspects, this article focuses on the applied logics of the software and discusses the approach in light of the dictionary's microstructure. The article concludes with some considerations about the clusters' advantages and limitations.

Keywords Historical lexicography; dictionary; word history; visualization; digital humanities; graph theory

1. Introduction: some characteristics of *Wortgeschichte digital*

Wortgeschichte digital (Digital Word History, or hereafter, WGd)[1] is a digital, monolingual dictionary of the German language that aims to describe the German vocabulary from about 1600 through today, focusing primarily on semantic shifts. A hallmark of the project is the narrative form of its entries, which describe, reflect, and historically contextualize observed developments. This distinguishes it from the majority of other dictionaries available. But WGd does in some respects follow the tradition of historical dictionaries nevertheless, enriching its entries, for example, with an appropriate amount of quotations to illustrate and provide evidence for historical usage. The project began from scratch in 2019; it is still in its early stages, and does not intend to describe the whole of German vocabulary (which would be beyond its personnel resources anyway). Instead, it aims to provide a selection of several thousand lemmas that belong to an array of different topic domains, such as *politics and society, economy,* and *music and arts.* As the dictionary explicitly addresses the public at large, its entries are written in a more relaxed style but without neglecting their scientific nature.

Given the broad target audience, the project strives to enhance user experience by implementing a host of technical tools to its website. Among the already available features are a timeline (*Zeitstrahl*) that serves as an alternative access point to the dictionary's lemmas; tightly integrated help windows that elucidate the usage and meaning of linguistic terms; a quotation navigator (*Belegnavigator*) that provides an overview of the quotations' chrono-

[1] https://wortgeschichten.zdl.org/.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.
https://doi.org/10.14618/phpy-6r66

logical distribution and the spread of historical word forms; and last but not least the cross-reference clusters (*Verweiscluster*) with a clickable, structured lemma list.[2]

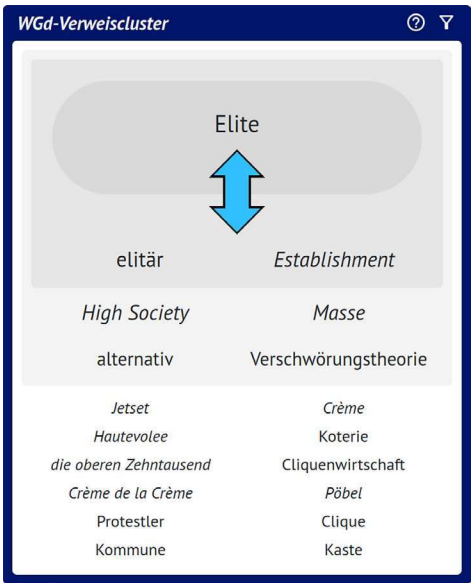

Fig. 1: A cross-reference cluster as displayed in the entry Elite; the lemmas are clickable and direct the user to the appropriate entry

This article provides a deeper insight into the cross-reference clusters. In section 2, it delivers an account of the rules devised to obtain a cluster (as shown in fig. 1). The section not only describes the applied rules but also discusses their underlying logic. Section 3 summarizes the advantages and limitations of this approach with a view toward other projects. First, though, we make some remarks on the microstructure of the WGd dictionary. The considerations in section 2 and the notes on the cross-reference clusters' limitations would otherwise not be comprehensible.

* * *

There are two core ideas or prerequisites on which the cross-reference clusters are based: First, they should allow users to find entries of interest within the dictionary. They should not have to rely on an alphabetical list as the sole entry point and search option. Therefore, from a visual perspective, principles of simplicity and perspicuity must be heeded. Second, the construction of the clusters must be fully automatic. This is of particular importance, as the WGd entries are published continuously rather than in installments, as is usually the case with print publications. That is why the form and content of each cluster are highly dynamic. Since the continuous publication of new entries constantly increases the complexity of the entangled web that links entries with one another, and since the clusters result from an analysis of that very network structure, they have to be redrawn every time a new entry hits the web. This creates a pressing situation, as the continuous publication also leads

[2] For the timeline, see https://www.zdl.org/wb/wortgeschichten/#Zeitstrahl. For the help windows, see, e. g., "Spezialisierung" in the first paragraph at the page https://www.zdl.org/wb/wortgeschichten/Masse. The quotation navigator can be found in every entry. Hit the navigator icon beside the heading "Belegauswahl."

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.
https://doi.org/10.14618/phpy-6r66

to the addition of references in already-published entries, especially when older entries mention a lemma that is the headword of a newly published entry.[3]

Harm (in this volume) provides an in-depth description of the entry's microstructure. But to grasp the conditions that mold the clusters, some key features of the WGd entries need to be outlined here as well.

Each WGd entry opens with a summary, followed by an orienting section. In addition to a table of contents, the orienting section incorporates some aspects that usually form the basis of a historical dictionary: listed word meanings and more or less extended word lists of up to three categories: word formations (*Wortbildungen*), word combinations (*Wortverbindungen*), and similar expressions (*bedeutungsverwandte Ausdrücke*). Especially the words listed in the latter category are enriched with details about semantic relations, such as contrast or synonymy. The orienting section precedes the core of every WGd entry: a continuous text that charts the semantic development that the headword underwent within the period under investigation. Many entries have initially hidden text passages that give users in-depth information about a discussed fact. A good example is in the entry *Erika Mustermann* ("Jane Doe"). Its so-called *Mehr erfahren* ("learn more") section outlines the adoption of new identity cards in Germany around 1980. In doing so, it provides some information as to why the authorities resorted to the word *Mustermann* ("average person," with the archaic connotation "wholesome person") for the cards' mock-ups that still circulate today.

In addition to standard entries, the dictionary consists of a small but growing number of overview articles that deal with a whole word field at once.[4] The microstructure of these entries does not vary significantly from the standard layout. They mainly differ in their much broader point of view on semantic developments, which entails some changes to the orienting section. In those entries, enumerated meanings do not make much sense as the text deals only with lemmas for which there are standard entries to provide a much deeper insight.

The situation is further complicated by the fact that the logic of "one entry, one lemma" does not apply to WGd. In many cases, on the contrary, it is of particular importance to describe the semantic change of several lemmas in a sole entry. A striking example is *Beaumonde/die schöne Welt,* where the latter lemma ("the beautiful world") is a loan translation of the first. Hence, the description of both is inextricably entwined. Entries of that type are quite frequent. Furthermore, entries may also have subordinate lemmas (*Nebenlemmata*) that are defined and described much like the main headwords, albeit less comprehensively. The entry *alternativ*, for example, also covers the multi-word unit *alternative Fakten* ("alternative facts"), an expression in which *alternativ* adopts the meaning "bogus, false".

References to different WGd entries can be found in every part of an entry, and they may point to either a main or a subordinate lemma. The cluster constructor evaluates all references regardless of their position in the entry, which can pose some issues.

[3] The lexicographers have a quality assurance tool (cf. section 2.2) that scans all entries once there is a newly published. It specifically tracks terms that are enclosed in special markup code (TEI `<mentioned>`) and external references to entries of our partner project (https://www.dwds.de/). A message is printed if the headword of a new entry matches one of those terms. The tool also ensures markup consistency, such as for diasystemic values and semantic relations. The *Svensk ordbok* relies on a somewhat comparable tool that visualizes cross-references to obliterate errors and gaps (Blensenius et al. 2021).

[4] E. g. https://www.zdl.org/wb/wortgeschichten/Wortfeld-Lebensformen.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.
https://doi.org/10.14618/phpy-6r66

Regarding multi-word entries, it is impossible to tell which lemma is exactly the referrer that points to a different lemma. In such a case, we have to resort to the assumption "all headwords point to the referred lemma". The underlying decision to write a multi-word entry is always based on the observation that the headwords need to be described in close conjunction to elucidate their semantic shift and/or historical distribution. We discuss together what belongs together. That is why it would be a difficult thing to partition such an entry into sections that deal solely with one of the lemmas. Besides, such a solution had to impose formal restrictions to the writing process that are not eligible in the light of the project's knowledge interest, even if they were preferable from a technical point of view.

Fortunately, the inverse case does not pose any problems. We can always be sure of which main lemma a reference points to. The same is true for references to subordinate lemmas, as those are always associated with a certain text position from where their description starts.

As mentioned before, some references have a semantic description attached to them. But that is not always the case, again given the specific structure and contents of WGd entries. It would be cumbersome and difficult to attach specific semantics to every single reference, as there are several reasons why they were added, and these reasons exceed baseline categories like synonymy or hypernymy by far: There are references that point to lemmas with a comparable semantic shift; others refer to lemmas that are part of the same word family; others point to the descriptions of a historical context that is given in a different entry and so on. Not all of these are linguistic categories. It would be quite difficult to tackle this issue, especially with limited personnel resources, but a denser markup would clearly have computational advantages (cf. Meyer/Müller-Spitzer 2010).

Again, all these issues arise mainly because WGd entries are written as a continuous text that includes information that exceed or differ largely from typical lexical resources. However, the approach chosen for the cluster constructor alleviates those issues substantially.

2. On constructing a cross-reference cluster

2.1 Basic rules

The construction of the cross-reference clusters is not as dependent on advanced programming skills as one might think. The usage of some basic logic and a limited understanding of a graph structure with directed edges is much more important to achieve the outcome as shown in figure 1. When the baseline situation is as intricate as outlined in section 1, a good first step is to prune every distracting factor. Therefore, as a first step, we temporarily leave aside all the difficulties that the complex structure of the WGd entries imposes on us. For the time being, the references will be stripped to the core, regarded as if no semantic relation were attached to them at all. Thus, they have no particular remarkable quality other than pointing. This implies of course that they all have the same weight, which means none has conspicuous importance, a condition that enables us to operate on them with a very basic logic that will lead to a clear, programmable solution.

When the quality of all references is identical, it is a valid approach to reduce the issue in a way that only three relational types remain:

1) A $\rightarrow$ B (A points to B)

2) B $\rightarrow$ A (B points to A)

3) A $\leftrightarrow$ B (A points to B *while* B points to A)

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.
https://doi.org/10.14618/phpy-6r66

The third type looks promising, especially when this relation pertains to more than two entities at once; in that case, the entities refer so densely to one another that they are virtually trapped in a reciprocal structure. That is an interesting signal of proximity. We cannot say anything about the quality of this proximity (as we stripped the references of the distinguishing semantics they might have had), but we are able to posit a first rule:

(1) Every time we encounter a reciprocal reference structure, all the involved lemmas belong to a cluster center (*Clusterzentrum*) when the lemmas are also distributed over at least two different entries.

The lemmas *Elite, elitär* ("elitist"), and *Establishment,* in the dark gray area of the example cluster in figure 1, abide by this rule because they exhibit a reference structure like this:

Elite ↔ *elitär*
Elite ↔ *Establishment*
elitär ↔ *Establishment*

In other words, every lemma of a cluster center points to every other lemma that belongs to the same center (otherwise they would not form a cluster center). What we see is a state of maximal reciprocity.[5] But if we had left it at that, we would have discarded a host of references because such a tight reference structure, as it can be observed within a cluster center, is the exception rather than the rule. Thus, to include at least some of the missed references we deploy another rule:

(2) Every time a lemma that belongs to a cluster center points to another lemma that does not belong to the same center, the latter lemma is part of the cluster fringe (*Clustersaum*).

In the example cluster of figure 1, the fringe encompasses all the lemmas in the light gray area, from *High Society* to *Verschwörungstheorie* ("conspiracy theory"). Again, the assumption behind the rule is that every reference is of equal importance. Therefore, when a lemma does not belong to a cluster center but is referred to from within that center, it will be closely related to it. That is even more the case if the particular lemma points back to a lemma within the center while that very lemma points to the fringe lemma. (If the fringe lemma had pointed to *all* the lemmas in the cluster center, it would have formed part of the center itself. But this can be ruled out since we checked that before.) This observation leads us to the next rule:

(3) Every time a lemma that belongs to the cluster fringe forms a reciprocal reference with a lemma within the cluster center, this lemma is more pertinent to the lemmas in the center than to the other lemmas at the fringe.

In the example cluster, this is the case for *High Society* and *Masse* ("mass") but not for *alternativ* and *Verschwörungstheorie*.[6] Thus, the first two are promoted and bunched into a distinctive bundle (*Bündel*), and as a consequence are printed above the latter lemmas in a larger font.

The next rule leads to the inclusion of all the lemmas against the white background, from *Jetset* to *Kaste* ("caste"). In our initial considerations, it follows the second relational type, at least on the assumption that entity A were a lemma of the cluster center and entity B a lem-

[5] This is a plain example because each lemma has its own entry. That is not always the case (see section 1).

[6] In fact, *High Society* points to *Establishment* and vice versa. The same is true for *Masse* and *Elite*. But this cannot be deduced from the cluster itself; one must consult the source code or the quality assurance tool (cf. fn. 3).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.
https://doi.org/10.14618/phpy-6r66

ma that did not pertain to that center (in that sense, the first relational type is already covered by rule 2):

(4) A lemma that points to a lemma of a cluster center belongs to the cluster periphery (*Clusterumfeld*) when no lemma of the cluster center points back to that very lemma from the periphery.

To be precise, the final clause of this rule is superfluous and only for clarity; for if a lemma from the cluster center points to another lemma, rule 2 is applicable, and that rule excludes a lemma automatically from being part of the cluster fringe because:

(5) A lemma must not appear in more than one of the three cluster circles (*Clusterkreise*).

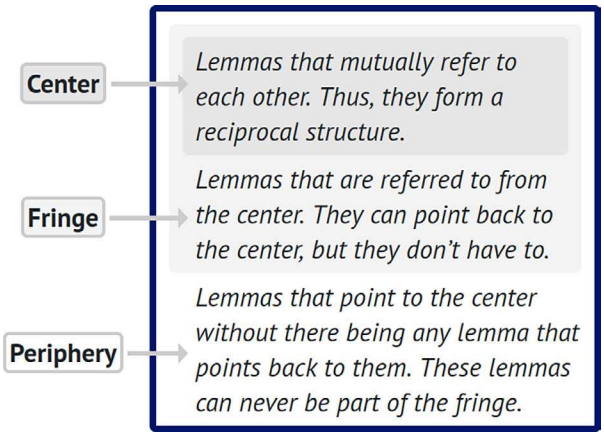

Fig. 2: The basic structure of a cross-reference cluster

What results from these considerations is the basic structure of a cross-reference cluster (fig. 2). This consists of three circles (center, fringe, periphery), whereas the fringe can be subdivided into two distinct bundles. It is important to reiterate that the resulting clusters are all constructed starting from the cluster center. From there on, the constructor follows its way down to fill in around the center. Thus, the starting point is an area in which the lemmas are closely related to one another. From there on, we proceed to include much more loosely affiliated lemmas. What all cluster lemmas have in common is a straightforward, direct connection to different lemmas in the center.

2.2 Adding further details

The next step is to enrich the clusters with further details we can obtain from the entries. For one thing, we know the position of a reference in the microstructure of a WGd entry. It is quite simple to recognize whether a reference is rather prominent (e. g., in the summary) or marginal (e. g. in a footnote). We also know about their degree of systematicity. A back reference that points to an overview article or the inclusion of a reference in a structured word list certainly signals that an author intends to systematize an observation. Those references should bear more weight than others. That is why we devised a point system, in order to attach weight to the lemmas and promote those that were referred to more often, more prominently, and/or with a greater degree of systematicity.[7] The points a lemma col-

[7] There are six different systematical positions where references can be found: The entry header may include a reference to a superordinate overview article (if there is one). For this is a quite prominent

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.
https://doi.org/10.14618/phpy-6r66

lects enable us to sort them by weight within their appropriate circle. What results are non-perspectivized clusters (i. e., clusters in their standard shape, see below), as they can be seen on the project's overview page (fig. 3).

Fig. 3: Two non-perspectivized clusters as they can be seen on the overview page

The word *Lebensformen* ("lifestyles") in the second cluster of figure 3 is printed in small capitals because it leads to an overview article that deals with the five lemmas that complete the cluster center. There is no specific rule that overview articles should be the first words of a cluster center, but the point system ensures that this is always the case. As every entry that pertains to a word field points back to the superordinate overview article, those references yield a lot of points to the field article, which promotes it in a way that it becomes the cluster's head.

position with a high degree of systematicity, such a reference yields 10 points. Thus, an overview article is always the sole head of its cluster. References in the word lists, which can be found in the orienting section, also exhibit high systematicity and therefore yield 3 p. Sometimes references can be found in the entry's summary (3 p.), a prominent position, but they are usually only part of the main text (2 p.). These 2 p. are something of a baseline on which the whole point system rests. Hence, references in initially hidden passages that provide further information are demoted, as are those in footnotes (1 p.). The subordinate lemmas pose a special problem, as there is no way of telling whether a reference comes from them or the article's main lemmas (see section 1). Therefore, it is assumed that an outgoing reference is always from the main lemmas. Fortunately, the inverse does not pose problems. We can reliably determine whether a reference points to a subordinate lemma so that the point system is not confounded by them. Finally, lemmas within the cluster fringe receive a 1000 p. bonus for every reciprocal reference they form with a lemma of the cluster center (see rule 3). This high number serves as a reliable marker, indicating that the lemmas pertain to the upper bundle of the fringe. Such a high value is never reached by references alone. The reason we initially applied the point system was to ensure that the superordinate overview articles are always the first lemma in a non-perspectivized cluster; thus the high score for references that point to them. Although the point system is a bit arbitrary, it ensures that lemmas that are referred to more often, more systematically, and more prominently are ranked higher than others.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.
https://doi.org/10.14618/phpy-6r66

Now that we can gauge the importance of a word in a cluster circle, we can formulate a threshold condition that splits the cluster center into separate bundles similar to the cluster fringe:

(6) When some lemmas of a cluster center gain significantly more points than the rest,[8] they are more pertinent and therefore promoted.

Thus, those lemmas are printed in the first position, with a gap below and in a larger font. This is the case for *Lebensformen* and *Elite* in figure 3.

The division of the cluster center into separate bundles is dismissed when a cluster gets perspectivized (as in fig. 1). In comparing the clusters in figure 1 and figure 3, you will notice that they are virtually the same. In actuality, they are the same because the two visualizations rely on the same data.[9] The only difference is that the entry lemma has been printed into an ellipse and a two-sided arrow has been added to indicate the reciprocal relation of the lemmas within the cluster center.

But there is another remarkable change. As the clusters in the entries are perspectivized, we are able to add back the semantic relations we initially dropped (see section 2.1). This step is possible only in a perspectivized cluster, as lexical relations like hypernymy and homonymy, meronymy and holonymy require a certain perspective in the form of a reference lemma to be valid declarations. Every cluster lemma printed in italics has a lexical relation attached to it. These are visible as a tooltip when the mouse hovers over them. Alternatively, the lemmas can be filtered by relation when one clicks the filter icon in the upper right corner (fig. 4).

Fig. 4: A filtered cluster that hides every lemma unless it is a synonym to the headword

[8] The threshold is derived from experience. While arbitrary, it works out well. It amounts to ≥ 10 % of the first lemma's points but at least 3 p. which is the score a prominent reference yields. As it is applied while the clusters are visualized, it is not present in the data itself.

[9] The data file is publicly available: https://www.zdl.org/wb/wgd/api#Artikeldaten.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.
https://doi.org/10.14618/phpy-6r66

2.3 Lemmas with multiple assignments

As said before, the clusters that result from the application of the given rules are based on the entire cross-reference network that pervades the WGd entries. That is why their default shape is a non-perspectivized cluster. Hence, the clusters are derived not from a word perspective but from the word fields our lexicographers are working on. During the writing process, we do not follow the alphabet but operate within thematic boundaries. That is a condition for the cluster constructor's analyses to produce meaningful results even if there is a relatively small number of published lemmas.

What results from the fact that clusters are based on word fields is that the lemmas have no special allegiance to a particular cluster. They can appear in more than one at the same time, even lemmas that are part of a cluster center. In this respect, a certain fuzziness is observed across the clusters. As we have seen, *Elite, Establishment,* and *elitär* form a cluster center, but *Elite* and *Masse* form a thematically adjacent, alternative center, as the two words exhibit a reciprocal reference structure as well. We do not take that to be an issue. Despite the multiple assignment of lemmas to several clusters the 187 lemmas that have been published to date form only thirty clusters.

Nevertheless, the lexicographers have access to a quality assurance tool that shows them a way to reduce the number of clusters (in addition to coding errors and suggestions for further references when applicable). This tool compares the clusters, lists them, and displays their degree of similarity. If a lexicographer deems it worthy to merge two or more clusters, the tool tells him or her which references need to be added to an entry in order to accomplish the goal.

Merging smaller clusters into larger ones is the only viable option. That is because the underlying program executes its rules thoroughly, which makes it very hard to mold a cluster into a shape of one's preference. We do not think that this poses a problem, for it was the lexicographer in the first place who decided to interlink two entries by adding a reference. The computer only visualizes the results of that decision and helps to keep entries in good shape. It also points to similarities that might have been overlooked. Yet the decision of whether a reference should be added, is not conferred to a computational device — and should not be. Nevertheless, the computers' ability to visualize data is indispensable given that we are dealing with an overload of information (cf. Therón et al. 2014). Visualizations like the cross-reference clusters will definitely alleviate this situation, as one of their strengths is rating of pertinent information.

A graph of the example cluster with directed edges not only illustrates the intricacy of the reference network that pervades the WGd articles (fig. 5) but also clarifies that, even if the number of vertices is low, the entanglement of the cross-references is at best hard to understand, even when the vertices are color-coded. Nevertheless, such a graph can be helpful, as it highlights the earlier-mentioned fact that lemmas may appear in more than one cluster center. The dark-blue-rimmed lemmas form the center of the example cluster in figure 1 (*Elite, elitär, Establishment*). Additionally, nodes 1 and 5 (*Elite, Masse*), nodes 3 and 4 (*Establishment, High Society*), and nodes 4, 8, and 9 (*High Society, Jetset, Crème*) form different centers, as their lemmas also refer to one another reciprocally. Furthermore, it is clear that the nodes 2, 5, 6, and 7 (*elitär, Masse, alternativ, Verschwörungstheorie*) are not part of the cluster formed by nodes 4, 8, and 9, as there is no edge that directly connects those lemmas.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.
https://doi.org/10.14618/phpy-6r66

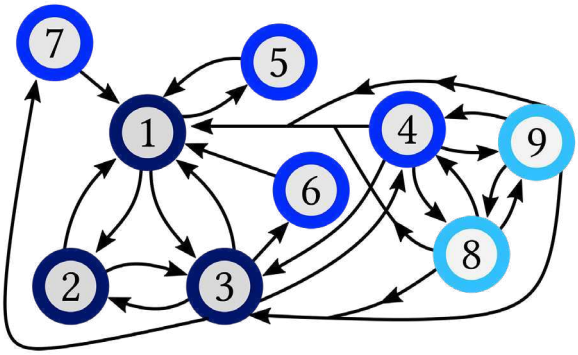

Fig. 5: A directed graph that displays all the references between the first 9 lemmas of figure 1, whereby vertex 1 is *Elite* and vertex 9 *Crème*

Figure 5 also points to the fact that the construction of cross-reference clusters is essentially a mathematical issue that pertains to graph theory. In that respect, we can reformulate the problem of finding cluster centers as a variant of a clique problem. What the cluster constructor actually does is find all cliques in a graph (like the one shown in fig. 5), whereas a clique is a subgraph that is complete in the sense that all vertices of the subset are adjacent. That means that all vertices have edges with every other vertex of the same clique. In graph theory a clique is also present if there is one edge that connects two different vertices. But the cluster constructor tries to find the maximal clique that is the largest subgraph that fulfills the requirements of a clique. Hence, smaller cliques that are subsets of larger cliques are discarded. Although the term *clique* is usually reserved for undirected graphs, that should elucidate the problem from a mathematical point of view. In contrast to a classic clique problem, what differs in the case of the WGd clusters is that the underlying graphs are directed and a clique is accepted as such only when all vertices have inbound *and* outbound edges to every other member of the clique.

3. Discussion: advantages and limitations

The cross-reference clusters are a perfect match for the dictionary's layout. They reflect its workflow, which does not follow the alphabet, but focuses on topic domains, word fields, and word families. Against this background, it is no surprise that the rules, as stated in section 2, reveal a host of reciprocal reference structures even in the project's early stages. Therefore, the technical solution (i. e. the cross-reference clusters) is very much in line with the project design and with the microstructure of its entries.

We believe that the outcome also satisfies the prerequisites as formulated in section 1. The clusters' shape is clear and subtle; it guides viewers by highlighting and promoting those lemmas that are more pertinent than others in a given context. Especially their perspectivized form, when the lemmas are enriched with details about semantic relations, seems to have much value. But we can still only surmise that this is the case. A comprehensive study, which tracks the way users interact with the WGd website, would be highly desirable.

We have already hinted at some of the limitations to this approach: The structure of WGd entries, which is quite loose for a dictionary, imposes some technical issues. But as the clusters are clearly an addendum, whereas the text is the gist of an entry, we do not intend to change much in that respect. Although the analysis could gain a bit from a more in-depth markup of every reference's relational meaning, we will probably never force multi-word

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.
https://doi.org/10.14618/phpy-6r66

entries to allot separate parts of the entry to one lemma alone. That would clearly contradict the project's mission and knowledge interest.

Of more concern, though, is that the cross-reference clusters don't scale well. It is hard to imagine a cluster with dozens of lemmas at its center. In that case, thresholds can reduce their size. We also are aware that the multitude of clusters that arise from this approach will gradually become overwhelming. Therefore, we already restricted analysis to topic boundaries. That means that, in practice, we don't take into account the whole reference network but subdivide it into thematic chunks. That does not exclude lemmas from, say, the topic domain *economy* from appearing in a cluster that was calculated for lemmas that pertain to *politics and society* (which is possible). But we do exclude lemmas from foreign topic domains during the detection of cluster centers.

This leads to another limitation. The calculation of cluster centers can be quite expensive in terms of computational workload — a well-known issue pertaining to the calculation of cliques. The current approach relies on analysis of all possible centers or cliques that can be derived from the references to a lemma. For example, there are currently five different lemmas that point to *elitär*: *Clique, Cliquenwirtschaft, Elite, Establishment,* and *Koterie.* In theory, every conceivable combination of these six lemmas (*elitär* has to be included) can form a cluster center. Therefore, possible combinations are as follows:

1) elitär, Clique, Cliquenwirtschaft, Elite, Establishment, Koterie
2) elitär, Clique, Cliquenwirtschaft, Elite, Establishment
3) elitär, Clique, Cliquenwirtschaft, Elite, Koterie

…

57) Establishment, Koterie

That should give an idea of what this actually means, as the number of combinations is already 57 in this example.[10] When we have to deal with six referring entries, the number rises

[10] There are 15 unique combinations with 2 lemmas, 20 with 3, 15 with 4, 6 with 5, and 1 with 6. A self-contained sample code that fills an array with all imaginable combinations looks as follows (in this case written in JavaScript):

```
let comb = [
  [
    [
     "elitär", "Clique", "Cliquenwirtschaft", "Elite", "Establishment", "Koterie",
    ],
  ],
];
let lemmas = 0;
let currentComb = [];
function makeComb (len, start) {
  if (len === 0) {
    comb[comb.length - 1].push([...currentComb]);
    return;
  }
  for (let i = start; i <= comb[0][0].length - len; i++) {
    currentComb[lemmas - len] = comb[0][0][i];
    makeComb(len - 1, i + 1);
  }
}
for (let i = comb[0][0].length - 1; i >= 2; i--) {
  comb.push([]);
  lemmas = i;
  currentComb = [];
  makeComb(i, 0);
}
```


This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.
https://doi.org/10.14618/phpy-6r66

to 120. One can easily imagine that this quickly leads to staggering figures and a huge computational workload. There are some tricks, though, to alleviate that problem. The 57 combinations in this example are not actually checked, as 53 of them can be discarded quickly. Consequently, this limitation does not pose a severe problem at the moment. But if one had to master much larger numbers of references and entries, it definitely would.

WGd is clearly not the first project that offers users visualizations with a navigational purpose. One might think of the word clouds in DWDS entries (*DWDS-Wortprofil*),[11] or the word graph in Wortschatz Leipzig.[12] At a first glance, these visualizations seem akin to the WGd clusters, as they are both an informational tool and a navigational device for users. But the underlying analyses are very different in terms of source data. Where WGd deals with manually set references by lexicographers, those projects operate with huge data sets and reveal cooccurrences.

On the contrary, the Semagrams of the Algemeen Nederlands Woordenboek (General Dutch Dictionary, ANW) have a lot to do with semantic relations. They are a good example of how a tight lexical structure can alleviate entries' findability. The ANW's Semagrams are highly structured lists of meanings that fill a predefined number of slots (Moerdijk et al. 2008). Each slot represents a semantic class and gives short informational descriptions like "size: is big", "place: is kept on a farm" for the entry *koe* ("cow").[13] Additionally, nonvisible data fields store keywords, synonyms and "relevant words" (ibid., p. 20). These are used for advanced database queries. But the ANW's Semagrams stem from a completely different lexicographical approach and are therefore not applicable to WGd.

A bit more in line with the style of WGd is *elexiko*.[14] The dictionary also deploys different entry types, including word group articles (*Wortgruppenartikel*) and multi-word entries that deal with sense-related lemmas (*sinnrelationale Paare und Gruppen*). Some of those entries include raster graphics that visualize semantic relations in the form of complex stemmas or intersected word fields. However, it seems that these graphics have not been calculated, but individually created for the specific purpose of an article in question. Contrary to WGd, *elexiko* uses a tighter XML structure for the cross-reference markup (cf. Meyer/Müller-Spitzer 2010). Therefore, an automatic visualization of the semantic relations should not pose much difficulty. With the exception of some similarities to the WGd macrostructure, the standard entries of *elexiko* rely much less on continuous text. Instead, they offer an astounding multitude of word information. In that sense, the dictionary is much more in line with the ANW.

In short: None of these projects drew on the same solution. That is certainly not a stunning observation, as all these tools and visualizations were devised on a different lexicographical basis and out of diverging knowledge interests. Eventually, it seems that every project needs a solution of its own that matches both its scientific approach and its dictionary's microstructure.

[11] https://www.dwds.de/d/ressources#wortprofil.

[12] https://wortschatz.uni-leipzig.de/en.

[13] https://anw.ivdnt.org/article/koe.

[14] https://www.owid.de/docs/elex/start.jsp.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.
https://doi.org/10.14618/phpy-6r66

References

Blensenius, K. et al. (2021): Finding gaps in semantic descriptions: Visualisation of the cross-reference Network in a Swedish monolingual dictionary. In: Proceedings of the eLex 2021 Conference, Brno, pp. 247–258. https://elex.link/elex2021/proceedings-download/ (last access: 25-03-2022).

Meyer, P./Müller-Spitzer, C. (2010): Consistency of sense relations in a lexicographic context. In: Proceedings of the Workshop "Semantic Relations. Theory and Applications". Malta, pp. 37–46. http://www.lrec-conf.org/proceedings/lrec2010/workshops/W9.pdf (last access: 24-03-2022).

Moerdijk, F. et al. (2008): Accessing the ANW dictionary. In: Coling 2008: Proceedings of the workshop on Cogintive Aspects of the Lexicon (CogALex 2008), Manchester, pp. 18–24.

Therón, R. et al. (2014): Highly interactive and natural user interfaces: enabling visual analysis in historical lexicography. In: Proceedings of DATeCH 2014. Madrid, pp. 153–158. https://dl.acm.org/doi/proceedings/10.1145/2595188 (last access: 24-03-2022).

Contact information

Nico Dorn
Akademie der Wissenschaften zu Göttingen
ndorn@gwdg.de

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.
https://doi.org/10.14618/phpy-6r66