Analysing Changes in Official Use of the Design Concept Using SweCLARIN Resources

Lars Ahrenberg, Daniel Holmer, Stefan Holmlid, Arne Jönsson
Department of Computer and Information Science
Linköping University, Linköping, Sweden
firstname.lastname@liu.se

Abstract

We investigate changes in the use of four Swedish words from the fields of design and architecture. It has been suggested that their meanings have been blurred, especially in governmental reports and policy documents, so that distinctions between them that are important to stakeholders in the respective fields are lost. Specifically, we compare usage in two governmental public reports on design, one from 1999 and the other from 2015, and additionally in opinion responses to the 2015 report. Our approach is to contextualise occurrences of the words in different representations of the texts using word embeddings, topic modelling and sentiment analysis. Tools and language resources developed within the SweCLARIN infrastructure have been crucial for the implementation of the study.

1 Introduction

What is the relation between architecture and design? Should they be seen as concepts in the minds of speakers or as professions where stakeholders sometimes compete and sometimes join forces to achieve their goals? In this paper, we try to answer such questions using the resources developed for the analysis of Swedish by Språkbanken Text and distributed through the SweCLARIN portal. More specifically we want to study whether there is a change in the denotations and connotations of four related words: *arkitektur*, 'architecture', *form*, 'form', *formgivning* (cf. German *Formgebung*) and *design*. In particular, are there changes in their use and, perhaps, signs of a convergence? The study is limited to Sweden and Swedish as spoken in Sweden. Its rationale is a hypothesis from colleagues working in design that there has been an increased effort to place architecture and design under the same umbrella, not least from the side of the Swedish government, and that this development has been detrimental for the design field.

Dictionary definitions of the four words vary. According to one of them[1], the word *design* was first observed in Swedish in 1948. It is defined there as *konstnärlig formgivning*, 'artistic form giving' using the older term *formgivning*. It is not found in the Historical Swedish Dictionary, SAOB, as the words for the letter D were compiled and published at the beginning of the 20th century. Over the years *design* has established itself as a synonym of *formgivning*, and also, as will be shown, become the more frequently used of the two. The word *arkitektur*, 'architecture', on the other hand, is defined as a scientific discipline with a related concrete meaning as 'artistic and technical design of buildings'. Thus, *arkitektur* can be defined in terms of *design* but also in other terms. Nowadays, also *design* can be studied at universities as a separate subject of study.

The word *form* has many meanings, one of them being 'artistic form'. It is used in this sense by the private organisation *Svensk Form*, 'Swedish Form', established in 1845, and its journal, simply named *Form*. The mission of this organisation is to stimulate or inspire good design, and it aims to attract individuals and companies that work within the areas of "form, design and architecture".

The main Swe-Clarin resources made use of in this study are the Sparv text analysis pipeline[2] (Borin

[1] Nationalencyklopedins Ordbok,'The National Encyclopedia dictionary', 1995 edition.

[2] https://spraakbanken.gu.se/sparv/#/sparv-pipeline

Lars Ahrenberg, Daniel Holmer, Stefan Holmlid and Arne Jönsson 2023. Analysing Changes in Official Use of the Design Concept Using SweCLARIN Resources. *Selected papers from the CLARIN Annual Conference 2022. Ed. by Tomaž Erjavec and Maria Eskevich.* Linköping Electronic Conference Proceedings 198, pp. 1–11. DOI: https://doi.org/10.3384/ecp198001

et al., 2016), the SenSALDO[3] sentiment lexicon (Rouces et al., 2019), and the Swedish Culturomics Gigaword Corpus (Rødven Eide et al., 2016). In addition, we have used the Gensim framework (Řehůřek and Sojka, 2010) for word embeddings and topic modelling, and the VADER tools (Hutto and Gilbert, 2014) for sentiment analysis.

The paper is structured as follows. Section 2 gives a short background on the recent political developments and decisions relating to the fields of architecture and design. Section 3 provides an overview of our data and presents our general approach. In section 4 we present our results and, finally, in section 5 we state our conclusions.

2 Historical Background

In 1997 the Swedish government proposed an action program for an area identified as *arkitektur och formgivning*. Two years later an official governmental report (SOU), entitled *Mötesplats för form och design*, 'A meeting-place for form and design'[4] proposed a new initiative for design. Although the focus was on design the report argued that it was reasonable that there was a clear connection to architecture, as its proposals related to buildings and building sites and the main events were to happen in the upcoming 'Year of Architecture' referring to the year 2001.

A few years later, in 2009, the Swedish Museum of Architecture was given a new responsibility to cover also "other fields of design" and its name was changed to ArkDes: The Swedish Centre for Architecture and Design. Its mission is "to increase knowledge of and cultivate debate about how architecture and design affect our lives as citizens"[5]. The name ArkDes is derived from the two words *arkitektur* and *design* using the first three letters from each word. More recently a new SOU-report was requested which was ready in 2015. With the title *Gestaltad livsmiljö: en ny politik för arkitektur, form och design*, 'Shaped habitat: a new policy for architecture, form and design', it brought the three concepts architecture, form, and design closer together and proposed the establishment for a new public body dealing with them jointly. However, this has not yet been realised.

There are some qualitative studies of design policies, often comparing national contexts, but we are not aware of any study of a national design policy that makes use of similar computational text analytical models.

3 Method and Data

Ultimately, a close reading will be required to investigate how a certain concept such as architecture or design has been framed and understood in a set of documents. Distant reading of the kind we perform here can be useful, however, to catch general patterns in usage and provide quantitative estimates of them. We study the selected terms by term frequency counts and by comparing their local contexts. Contexts can be modelled in different ways each of which amplifies a different aspect or layer of meaning. We make use of three common models of context: word embeddings, topics, and sentiments. To generate these models we make use of open sourced software that is available on the GitHub platform. SweCLARIN resources are used for pre-processing, notably for parsing Swedish text and for supplying necessary lexical data on sentiments of Swedish words.

In addition to the two SOU reports mentioned above we have used the news sections of the Swedish Culturomics Gigaword Corpus (Rødven Eide et al., 2016) from relevant time periods for comparisons. Also included are 229 responses[6] to the SOU report from 2015 to see to what extent they use the relevant terms in the same way as the report itself.

Responses expressing opinions on the proposals and general contents of SOU reports can be submitted by anyone. However, usually, a number of organisations with an interest in the subject matter of the report are invited to do so. The responding organisations can be categorised according to different criteria: being

[3] https://spraakbanken.gu.se/resurser/sensaldo

[4] SOU 1999:123

[5] https://arkdes.se/en/about-us/

[6] All SOUs and responses included in our analysis can be found at https://github.com/holmad/Analysing-Changes-in-Official-Use-of-the-Design-Concept-Using-SweCLARIN-Resources

from a Sector such as Private or Public (8 different), having a different Legal Status such as a Company, Interest organisation, Consultants, Municipality, Museum etc (17 different), or according to Area of Interest, such as Architecture, Research, Urban construction (18 different).

In the course of the study, we discovered that the terms of interest were often joined as conjuncts of a coordinate structure such as *arkitektur, form och design*, architecture, form, and design, or *arkitektur, formgivning och design*. In fact, in SOU 2015:88 it turned out that more than 50% of the instances of the words *arkitektur* and *design* occur as part of this coordination. For this reason, we decided to look at it as a concept of its own and created versions of the report where it is handled as a single token. In the sequel we will refer to it as the triad and the modified corpus will be referred to as the retokenized version, see Table 2.

Variant type	**Examples**
Base form	arkitektur, form och design
Uppercase first letter(s)	Arkitektur, form och design Arkitektur, Form och Design
Hyphenation	arkitektur, form och de-sign arkitektur-, form- och design
Misspellings	arkitektur, from och design
Misread PDF file	arkitektur, form och design,dnr
Definite variant + genitive	arkitekturen, formen och designen arkitekturens, formens och designens
Synonyms	arkitektur, formgivning och design arkitekturens, formens och gestaltningens
Compounds	arkitektur-, form- och designpolitiken arkitektur-, form- och designfrågor arkitektur-, form- och designområdena,

Table 1: Variants of the triad.

The triad actually appears in many different variants in the documents. Table 1 shows some of the most common variants. Some of these variants are mainly orthographic: the use of uppercase or lowercase, first letter of the first or all nouns, the inclusion of a hyphen in one of the words, or enclosing of it in quotation marks and other punctuation marks. Other variants are morphological using definite and/or genitive forms instead of the indefinite, nominative form. All of these could be caught by a regular expression. The triad may also be embedded in a compound where the second part is a noun such as *politik*, ‘politics’ and *område*, ‘area’. We decided to treat all orthographic and morphological variants as instances of the triad, excluding those where it is part of a compound. These are quite numerous, however, supporting the view that the triad has developed into a conceptual unit of its own.

Frequencies for the terms of interest in the different datasets are shown in Table 2. For the SOU reports and the responses to the 2015 report figures are given for both the original and the retokenized versions.

Corpus	**arkitektur**	**design**	**formgivning**	**form**	**arkitektur, form och design**	**all tokens**
Gw 1990-99 (News)	793	1,008	204	10,421	0	60,037,845
Gw 2010-2015 (News)	1,595	4,042	303	28,102	0	168,998,305
SOU 1999:123	43	358	139	180	0	42,263
- retokenized*	38	353	134	180	5	42,248
SOU 2015:88	318	328	26	301	0	34,739
- retokenized*	126	136	18	117	192	34,163
Responses to SOU 2015:88	2023	1659	29	1601	0	266,675
- retokenized*	1114	724	20	676	1311	262,742

*The triad *arkitektur, form och design* considered as one unit, see Section 4

Table 2: Frequencies of the investigated words in different corpora. The counts refer to tokens including orthographic and morphological variants.

As can be seen in Table 2 the two SOUs are roughly the same size, but the frequencies of the various words differ, as will be further discussed in Section 4. We also note that the total text of the responses to the 2015 SOU is much larger than the SOU itself.

4 Analyses

We have analysed the two SOUs and the 2015 Responses from four different perspectives. We first present results from a frequency count of the various terms, including the triad. We then present results considering the terms' use in general language and their semantic space by looking at word embeddings similarities. Finally we present results from topic analyses and sentiment analyses related to the terms.

4.1 Term Frequencies

The frequency of the terms is presented in Table 2, and illustrated in the bar charts of Figures 1 and 2.

A first observation is that the ratio of *design* to *formgivning* is changing rather rapidly. This is true for the news corpora where the ratio for the 1990s is about 5:1, rising to 13:1 for the period 2010-2015. The same holds for the SOUs where the ratio goes from about 2.5:1 in the 1999 SOU to close to 13:1 in 2015. Actually, the word *formgivning* is not only losing ground to *design* but also to *form*.

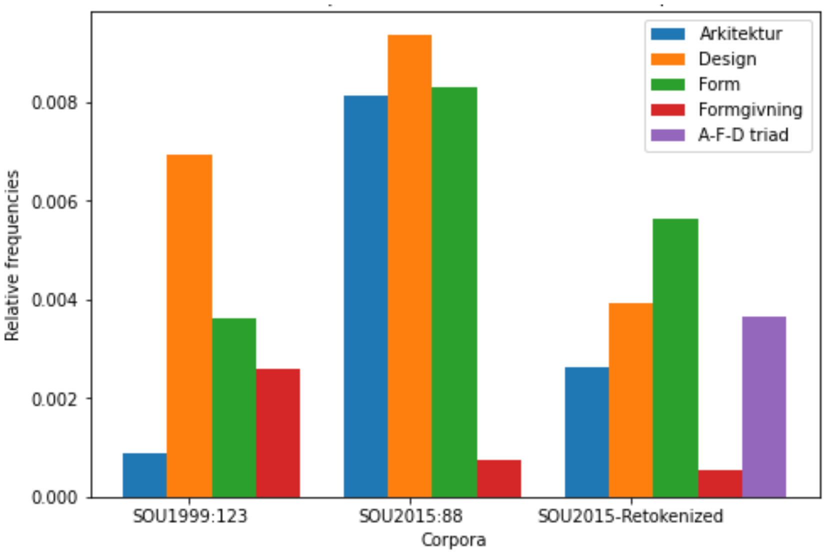

Figure 1: Relative term frequencies in the SOU reports.

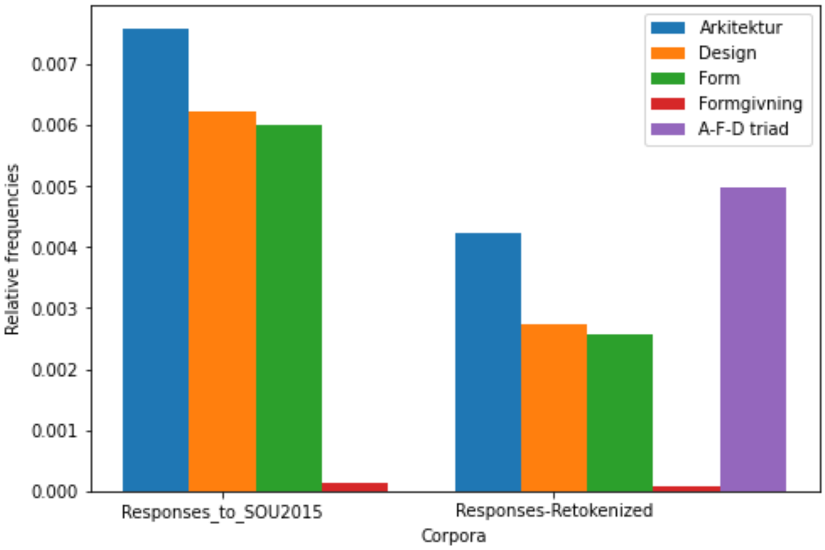

Figure 2: Effects of retokenization in the corpus of responses.

Secondly, we can see clearly the presence of the triad both in the report from 2015 and its responses. In fact, it is even more pronounced in the responses than in the report itself. See Figure 2.

4.2 Word Embeddings

To understand how the terms are used in general language, we looked at their distribution in the news sections of the Swedish Gigaword Corpus (Rødven Eide et al., 2016), for the periods 1990-99 and 2010-2015. Using several runs of word embeddings derived with the Gensim Word2Vec framework, we could observe the following:

- *design* and *formgivning* are close (synonyms) for both periods. The word *grafisk*, 'graphical', a common attribute to both terms, is about equally close.
- In the 1990s *formgivning* is a close neighbour to *arkitektur*, while *design* is further away. In the period 2010-15, the situation is reversed. In this period, *design* and *konst*, 'art' are competing for the place as the closest neighbour to *arkitektur*.
- The word *form* does not turn up in the close vicinity of any of the other words. This is due to its many other, more common meanings such as type, sort, shape, state, and mould.

Part of the reason why the words turn up as close neighbours in vector space is that they are often coordinated, in pairs such as *arkitektur och design*, but also in triples or even longer ones. Words that are frequent in these coordinations, apart from the four words under study, are *konst*, 'art', and *hantverk*, 'crafts'. We also see trends of concept building via these coordinations, for example in the name of the Museum of Architecture and Design, ArkDes. These events in general language correspond well with the developments we see in the government documents and especially in the SOU from 2015, where the triad is so frequent.

We also produced word embeddings for the two reports. Due to the smaller size and the random character of word embeddings, these are harder to interpret. An interesting observation, however, is that in embeddings generated from the retokenized versions the neighbourhood of the triad does not include the individual terms, and vice versa.

In order to compare the semantic space of the studied terms in the two reports, we used the temporal word analogies method as suggested by (Szymanski, 2017). This method works by transforming two vector space models into a common vector space, which acts as a link between the models, enabling the comparison of word vectors between two otherwise independent models. Thus, we can investigate shifts between the models, in the form of *"which word X in model M1 correspond to word Y in model M2?"*.

We trained a Word2Vec model for each of the reports and applied the temporal word analogies technique to search for differences in usage of the studied terms, *architecture, design*, and *form*. Although the models themselves showed some differences when extracting and manually inspecting their most similar words, this method did not reveal any semantic shift of any of the studied terms between the two reports. It is possible, however, that this is due to the relatively small size of the data and vocabularies used.

4.3 Topic Modelling

We have applied topic modelling to the reports to see whether they differ in their distribution of topics, using the Gensim package (Řehůřek and Sojka, 2010) on parsed versions of the reports. Gensim provides an implementation of Latent Dirichlet Allocation (LDA) (Blei et al., 2003), widely used for topic modelling applications. The number of topics per model was chosen to maximise the coherence score C_v (Röder et al., 2015), which resulted in the model for the 1999 SOU having 16 topics, and the 2015 SOU having 14 topics.

We split each SOU according to its chapter. Since the chapters are few but lengthy, we performed a second splitting process where each chapter was divided into chunks of $\approx$100 tokens. Each of these chunks was seen as a document in the topic modelling process. We tokenized, lemmatized and POS-tagged each document with the Sparv pipeline. This allowed us to only include words in their lemmatized form, as is fairly common practice, and to only include content words (nouns, adjectives, verbs, and adverbs), which should make the topics more interpretable.

We could see for the 2015 SOU, that for the topics where *design* is among the 10 most relevant terms, so is *arkitektur*, and vice versa. For the majority of topics where this happens, *form* is also among the 10 most relevant terms.

In addition to the topic modelling of the SOUs, we trained topic models also on the responses to the 2015 SOU. This was done to enable studies of the topic distributions on different categories of responding organisations. For this task, we used the BERTopic library (Grootendorst, 2022) that leverages the recent years' rise of pre-trained transformer-based language models and is able to produce topic models based on the semantic structure (rather than only the word frequencies) of a collection of documents. The library utilizes the pre-trained language model (in our case, we used a Swedish sentence-transformer model (Rekathati, 2021)) to create document embeddings, cluster the embeddings, and through a class-based TF-IDF procedure generate topics.

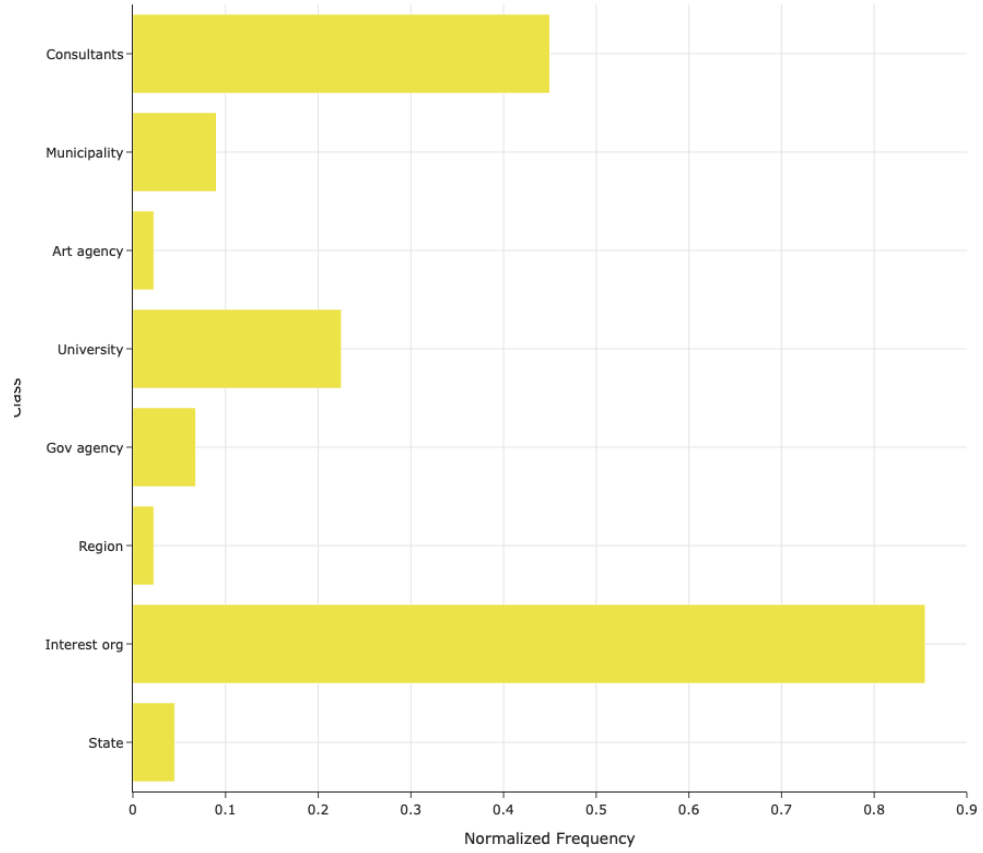

Figure 3: Relative frequency of a topic related to the term *design* across responses from different category classes.

The topic model was trained from all the SOU 2015 responses in our dataset. Within the BERTopic-framework we used the default UMAP model for dimensionality reduction and HDBSCAN for clustering (both with default parameters). The resulting model consists of 106 topics. Since all responses were assigned a category of origin, we could utilize the support BERTopic provide to visualize topics per predefined class.

In Figure 3, the topic most heavily related to the term design, including the word *design* itself and compounds such as *designmetodik*, 'design methodology', *designkompetens*, 'design competence', etc, are displayed to have the most prevalence in responses from respondents assigned to the legal status Interest organizations and Consultants. This design topic was highlighted only by organisations from 8 of the 17 legal statuses. From the wider range of actors, this topic did not emerge at all.

A topic heavily related to the term *arkitektur*, including words like *bostäder*, 'residences', *människor*, 'humans', and *förtätning*, 'urban consolidation', can be observed to just like *design* have a high presence in responses from Interest organizations as shown in Figure 4. This topic, however, is also often located in responses from Municipality, University, and Government Agency actors. Many categories of actors

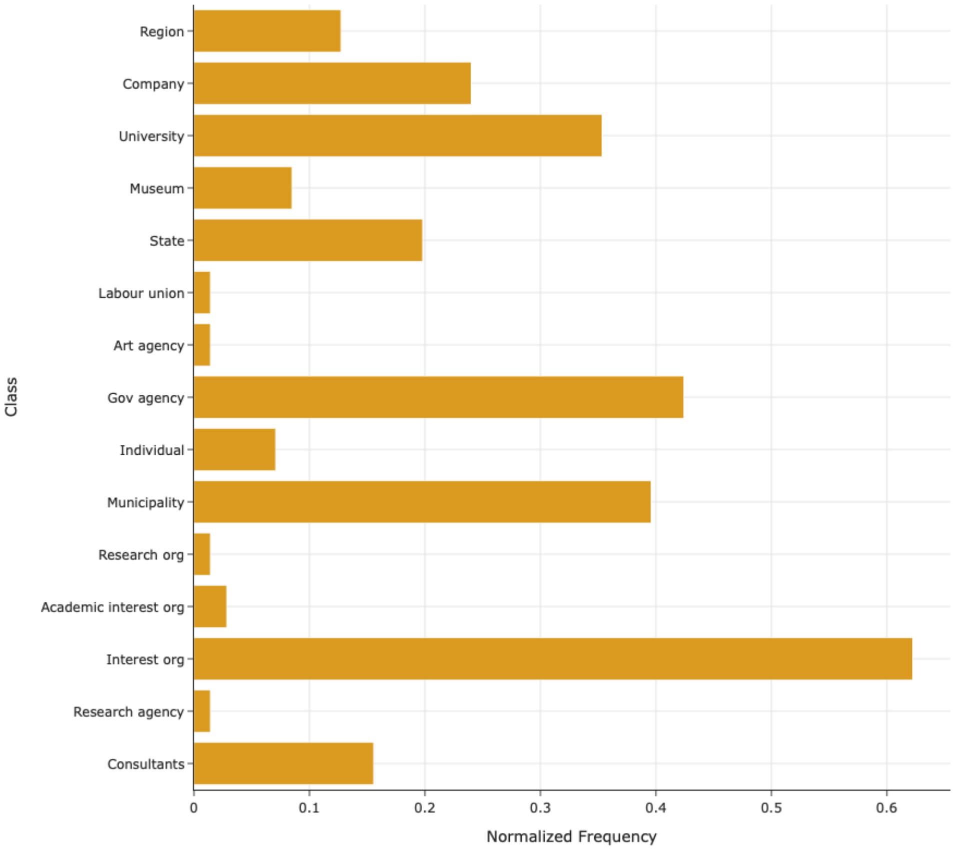

Figure 4: Relative frequency of an *arkitektur*, ’architecture’, related topic across responses from different category classes.

find it important to mention this topic in their responses, and trivial visual inspection shows that the ones highlighting it the most are the urban planning and commissioning actors as well as part of the businesses engaged. However, it is not possible from this analysis to see whether they agree with each other, whether they agree with the SOU or whether they disagree with the SOU.

4.4 Sentiment Analyses

Sentiment analysis was applied to give another perspective on the reports. In particular, we were interested to see whether there is a change in the way the relevant terms are presented in the two reports.

For sentiment analysis we used Vader[7] (Hutto and Gilbert, 2014) and the Swedish SenSALDO 0.2 sentiment lexicon (Rouces et al., 2019). The lexicon in English Vader comprises 5500 lexical entries with sentiment scores between +5 and -5. There is a Swedish version with a lexicon comprising 2067 lexical entries and sentiment scores +3 and -3, but less granular (Borg and Boldt, 2020). SensSaldo uses the three sentiment scores -1, 0 and +1. What makes SenSALDO unique in a Swedish context is that it assigns different sentiment values to different senses of a word, for instance, the Swedish word *fara* can mean ‘danger’ or ‘go (away)’ where the former has a negative sentiment and the latter is neutral. SenSALDO comprises 12287 lexical entries where 8893 are unique words. Word sense disambiguation with the SenSALDO 0.2 lexicon is made possible by using the Sparv pipeline. Vader also uses booster words, such as *amazingly*, to further refine the sentiment analysis. The booster dictionary used in the analyses is a slightly enhanced version of the dictionary used for sentiment analysis of e-mail conversations (Borg

[7] https://github.com/cjhutto/vaderSentiment

and Boldt, 2020) and comprises 89 items.

Vader produces a compound score for each sentence, by summing the valence scores of the words according to their identified sense and normalising this sum to be between -1 and +1. It is also useful to calculate the amount of positive, negative or neutral sentences. For this, we use the recommendations that a sentence has positive sentiment if the compound score is ≥ 0.05, neutral if the compound score is $>$ -0.05 and $<$ 0.05 and negative if it is ≤ -0.05[7].

Generally speaking, we find the sentiment score produced by Vader intuitively correct, although it has some problems with negations. A typical sentence with a positive sentiment from the responses, with a compound score of 0.8402 is: "*Med bättre kunskap och medvetenhet om vad god arkformdes innebär för människors välbefinnande kan vi skapa processer som långsiktigt främjar en god samhällsekonomi och goda livsmiljöer för alla*" 'With better knowledge and awareness of what good arkformdes means for people's well-being, we can create processes that promote a good social economy and good living environments for everyone in the long term'.

A negative sentence withe a compound score of -0.6124 is "*Kritiken har i hög grad skjutit in sig på brister i ledningen av verksamheten, men i själva verket har nog uppgiften att vidga tidigare Arkitektmuseets uppgifter till att även omfatta design varit olämplig*", 'The criticism has largely focused on shortcomings in the management of the business, but in fact the task of expanding the previous tasks of the Museum of Architecture to also include design has probably been inappropriate'.

And, finally, with sentiment score 0, a typical neutral sentence is "*Ett sådant exempel är de strukturer som förekommer i södra Sverige, med samarbete mellan bland andra Region Skåne, Malmö Stad, Form/Design Center, regionala branschplattformar, högskolor och universitet*", 'One such example is the structures that exist in southern Sweden, with collaboration between, among others, Region Skåne, Malmö City, Form/Design Center, regional industry platforms, colleges and universities'.

SOU	**Sentences**	**Negative Sentences**	**Positive Sentences**	**Neutral Sentences**	**Mean sentiment**
1999	2161	118 (6%)	820 (38%)	1223 (56%)	0.112^*
2015	1581	87 (6%)	722 (49%)	722 (45%)	0.168^*

Table 3: Descriptive statistics. *Significant, p $<$ 0.01

SOU	**Form**		**Formgivning**		**Arkitektur**		**Design**	
	Sentences	Sentiment	Sentences	Sentiment	Sentences	Sentiment	Sentences	Sentiment
1999	180	0.160^*	139	0.125	43	0.176^*	358	0.141^*
2015	301	0.275^*	26	0.254	318	0.288^*	328	0.248^*

Table 4: Concept-based sentiment for the SOUs from 1999 and 2015 in original versions. The number of sentences and mean concept sentence sentiment. *Significant, p $<$ 0.01

The results from analysing the two SOUs are depicted in Table 3. In both SOUs the amount of sentences having a negative sentiment is small, roughly 6% each year. The main difference is that the number of positive sentences is higher in 2015 than in 1999, and, correspondingly, the amount of neutral sentences is lower in 2015 compared to 1999. This can be seen in the mean sentiment. For the 1999 SOU, it is 0.112 and for 2015 it is 0.168. The difference is significant[8], p $<$ 0.01[9]. Thus, the 2015 SOU uses overall a more positive tone.

We have also investigated the sentiment for each of the concepts in focus *arkitektur, design, form*, and *formgivning*. For each concept, sentiment is computed if the concept occurs in the sentence. The results are presented in Table 4 showing that the text in the 2015 SOU has a more positive attitude towards the concepts *form*, *arkitektur* and *design*. The difference in sentiment for *form* between 1999 and 2015 is significant[10]. The difference in sentiment for *design* between 1999 and 2015 is significant[11], and the

[8] Welch's t-test = 7.2511, p = 0.0000.

[9] For all further significance tests we use p $<$ 0.01 to denote a significant difference.

[10] Welch's t-test= 4.8204, p = 0.000

[11] Welch's t-test = 5.7667, p = 0.000

difference for *arkitektur* is significant[12]. The difference for *formgivning* is not significant for p $<$ 0.01, but p = 0.0230 so there is a tendency.

SOU	Form		Formgivning		Arkitektur		Design		triad	
	Sents	Sentiment	Sents	Sentiment	Sents	Sentiment	Sents	Sentiment	Sents	Sentiment
1999	180	0.160	134	0.125	38	0.174	353	0.140	5	0.192
2015	117	0.239	18	0.172	126	0.266	136	0.170	192	0.304

Table 5: Concept-based sentiment for the SOUs from 1999 and 2015 in retokenized versions. The triad is a separate concept. The number of sentences and mean concept sentence sentiment.

If we consider the triad *arkitektur, form- och design* and use the retokenized corpus to filter out all sentences containing it, see Table 5 for descriptive statistics, there are no significant differences.

The triad is not used much in 1999, only 5 occurrences, so we also compared the triad to the other concepts for 2015, last row in Table 5, and then it turns out that only the difference in sentiment for *design*, 0.168, compared to the triad, 0.304, is significant, $p < 0.01$. That is, the triad is presented with a more positive sentiment than *design* in the 2015 SOU.

Corpus	Form		Formgivning		Arkitektur		Design	
	Sentences	Sentiment	Sentences	Sentiment	Sentences	Sentiment	Sentences	Sentiment
Retokenized	117	0.239	18	0.172	126	0.266	136	0.170*
Not retokenized	301	0.275	26	0.254	318	0.289	328	0.248*

Table 6: Concept-based sentiment for the SOU 2015 filtered for the triad or not. The number of sentences and mean concept sentence sentiment. *Significant, p $<$ 0.01

In Table 6 we compare the last rows of Table 4 and Table 5 without the triad column to see whether there are sentiment differences for the four concepts considered. Again the only significant difference is for *design*[13].

Corpus	Form		Formgivning		Arkitektur		Design		triad	
	Sents	Sentiment	Sents	Sentiment	Sents	Sentiment	Sents	Sentiment	Sents	Sentiment
Responses	676	0.171	20	0.071	1114	0.180*	724	0.142	1311	0.224*
SOU 2015	117	0.239	18	0.172	126	0.266*	136	0.170	192	0.304*

Table 7: Concept-based sentiment for the 2015 SOU and the responses with the triad as a separate concept. The number of sentences and mean concept sentence sentiment. *Significant, p $<$ 0.01

The sentiment analyses of the responses to the 2015 SOU are presented in Table 7. The difference for the concept *arkitektur* in the SOU, 0.266, and the responses, 0.180, is significant[14], and the difference for the triad SOU = 0.304, responses = 0.224, is significant[15], i.e. the concepts are used more positively in the SOUs compared to the responses.

5 Conclusions and Reflections

By using a variety of methods and SweCLARIN resources we were able to present textual analyses of the material from different perspectives. The tools and language resources available in the SweCLARIN infrastructure for analysis of Swedish texts enable comparisons of language use also over such short time spans as 20 years. In particular, we exploited the ability of the Swedish SenSALDO lexicon to identify word senses for sentiment analysis, and the Culturomics Gigaword Corpus for comparing official government reports with general language. Most of our analyses, the sentiment analysis, the topic modelling and the frequency calculations, utilized texts parsed using the SweCLARIN Sparv pipeline.

Although the SweCLARIN resources are in most cases straightforward to use, some fine-tuning of Sparv was needed. It would not have been possible to do the analyses without further programming, to

[12] Welch's t-test = 3.5494, p = 0.0007

[13] Welsh's t-test = 3.2559, p = 0.0013

[14] Welsh's t-test = 3.5261, p = 0.0006

[15] Welsch t-test = 3.7176, p = 0.0002

prepare texts for analysis using regular expressions, and to adopt libraries, such as Vader, to the language resources. For modelling of context we used open source software available on other platforms for sharing resources such as GitHub. Simplifying somewhat we can say that SweCLARIN supplied the necessary language-specific tools and lexical resources while the language-independent modelling software was obtained elsewhere. We think that researchers who want to apply language technology tools to their problems need to be aware of several repositories and what they can offer. Guides for their use, and particularly, when you need to combine them with language-specific data and pre-processing, can be part of pedagogical material and the handbooks provided by CLARIN centers. SweCLARIN, as an example, maintains such a handbook[16].

We have analysed and compared two public government reports, both being part of a political process of policy development, suggesting how financial and structural support should be developed for the areas architecture, design and form giving The first was published in 1999 and the second in 2015. We have also included responses to the 2015 report in the analyses. Sentiment analysis seems to be the method that provides the most reliable results in our case, while the results from topic modelling and temporal word analogies are more uncertain due to the small dataset. However, with the BERTopic library, we could relate topics to metadata of the responses.

On the one hand, we can see the word *design* being used with increased frequency overtaking the role of the older word *formgivning*. This reflects a general shift in language usage. On the other hand, when we look at the relation between the words *design* and *arkitektur*, we can see indications of semantic convergence, especially in the latter report. This is not least due to the creation of a vague super concept expressed by the coordinate structure *arkitektur, form and design*, which we refer to as the triad. In the former report, the triad is very uncommon. In the latter report either the triad or one of the individual terms is used, indicating that the suggestion of the report is to establish this super concept as a single area of policy. The triad, which is so frequent in the 2015 report, is significantly more positively described there, compared to *form* and *design* as individual terms, seemingly underlining a supporting policy that should take an integrative approach. Moreover, the triad is used in more positive terms than *design*, while the sentiment of the term *arkitektur* is not as influenced by the usage of the super concept. This indicates a possibility that there is bias in the report, or that there are more issues to deal with in the area of design than in architecture.

In the responses to the report from 2015, the triad is reused, and taken onboard as a concept to comment on. Sentiment analysis shows that the super concept is used more positively in the report than in the responses, indicating that there may be criticism in the responses toward the conflation of the three areas. Moreover, in the responses, few topics were related to design, and only a small number of actors were responding about design. This indicates that the majority of respondents were not reacting much to how *design* was handled in the report.

Overall, the analyses make it possible to identify potential aspects that could be further analysed for policy development. Structural issues, such as how certain areas are handled in relation to others, or what topics different clusters of actors are engaged in, can be identified. Also, content issues can be identified, such as how much focus different topics are given, but also more specifically how different interest areas are populated.

References

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Anton Borg and Martin Boldt. 2020. Using VADER sentiment and SVM for predicting customer response sentiment. *Expert Systems with Applications*, 162:113746.

Lars Borin, Markus Forsberg, Martin Hammarstedt, Dan Rosén, Roland Schäfer, and Anne Schumacher. 2016. Sparv: Språkbanken's corpus annotation pipeline infrastructure. In *SLTC 2016. The Sixth Swedish Language Technology Conference, Umeå University, 17-18 November, 2016*.

[16] https://sweclarin.se/swe/handbok

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Clayton J. Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI.*

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA. `http://is.muni.cz/publication/884893/en`.

Faton Rekathati. 2021. The KBLab blog: Introducing a Swedish sentence transformer. https://kb-labb.github.io/posts/2021-08-23-a-swedish-sentence-transformer/.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, page 399–408, New York, NY, USA. Association for Computing Machinery.

Jacobo Rouces, Nina Tahmasebi, Lars Borin, and Stian Rødven Eide. 2019. Sensaldo: Creating a sentiment lexicon for Swedish. *LREC 2018 - 11th International Conference on Language Resources and Evaluation*, pages 4192–4198.

Stian Rødven Eide, Nina Tahmasebi, and Lars Borin. 2016. The Swedish Culturomics Gigaword Corpus: A one billion word Swedish reference dataset for NLP. In *Digital Humanities 2016. From Digitization to Knowledge 2016: Resources and Methods for Semantic Processing of Digital Works/Texts, Proceedings of the Workshop, Krakow, Poland.*

Terrence Szymanski. 2017. Temporal word analogies: Identifying lexical replacement with diachronic word embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 448–453, Vancouver, Canada, July. Association for Computational Linguistics.