Exploring corpus linguistics approaches in linguistic landscape research with automatic text recognition software

Peter Gilles / University of Luxembourg and Evelyn Ziegler / University of Duisburg-Essen

Abstract
Taking a more quantitative approach in linguistic landscape research, we explore recent techniques of automatic information extraction from images. The recently released Cloud Vision API by Google offers new perspectives on the software-assisted processing and classification of pictures. A software interface makes it possible to extract various kinds of information from pictures automatically, among them the written text, certain labels to describe the picture (e.g. road sign, shop sign, prohibition sign) or the colours used in the picture. Applying this new technique to large-scale image data collections will not only enhance analysis but may also offer hitherto unrecognized structures.
The data comes from a large-scale investigation of the Ruhr Metropolis in Germany, where 25,504 photos have been taken to document the linguistic landscape for selected neighbourhoods in four cities (Ziegler et al. 2018). This data has been annotated manually in various categories to analyze the occurrence, form and function of visual multilingualism. These pictures are then automatically processed by the Cloud Vision API and the results compared to the manual annotation. It will show that the quality of the image recognition greatly depends on the quality of the picture. The textual information extracted from the pictures will be stored in a database.
Rather than presenting results on the linguistic landscape, this chapter is predominantly concerned with practical tools to facilitate large-scale linguistic landscape research.

Keywords: methodology, corpus linguistics, automatic text extraction, text identification

1. Introduction

Over the last 20 years, various methods have been established in linguistic landscape research in order to collect and analyze data. According to Androutsoupoulos (2014: 82–88), Van Mensel et al. (2016: 426f) and Pütz/Mundt (2019: 5–7), the following approaches to exploring the formation and meaning of linguistic landscapes can be distinguished:

- A predominantly quantitative, distributive approach characterized by counting signs and languages, mapping their (co-)occurrence and measuring linguistic diversity in a given area (cf. Peukert 2013, Gaiser/Matras 2016, Buchstaller/Alvanides 2017). According to this approach, language on public display is a symbolic means of place-making, indexing ethnicity and the status and prestige of languages, i.e. language regimes. In this perspective, place-making becomes an explanatory construct that is used to arrive at more general statements about ethnolinguistic vitality and the power relations between different linguistic groups.
- A more qualitative set of approaches which grew out of the concern that in the quantitative paradigm, social meaning is primarily assigned to signs and languages from an etic, not emic, perspective. Central to qualitative approaches is their focus on the relation between sign, place and addressee. This can take different forms. Following Scollon/Scollon (2003), a geosemiotic perspective has prevailed that takes into account the placement of signs and – in expanding the concept of interaction – the interaction between signs and addressees (Auer 2009, Domke 2014). There is also a growing body of research that is interested in (a) the historical conditions under which the textification of space has developed (cf. Blommaert 2013, Gilles/Ziegler 2019) and (b) the

ideological beliefs in which language choices and their perception are grounded (Garvin 2010, Ziegler et al. 2019).

Our study is situated in the quantitative paradigm. We will expand the use of quantitative data analysis methods by introducing a corpus linguistics approach to identify the dominant micro-linguistic characteristics shaping a given linguistic landscape. Previous research focusing on linguistic aspects has tended to be limited to (i) questions of language choice, (ii) practices of translation (Reh 2004, Sebba 2013) and (iii) fine-grained descriptions of the linguistic forms and structures found in text on signs (for German signage see Auer 2010, Hennig 2010, Schmitz 2017, Ziegler et al. 2018), while a large-scale quantitative textual analysis has not been undertaken to date.

We address this gap by presenting a new approach to written data on signs using a text identification software tool. The proposed method of automated text identification and extraction offers an opportunity to effectively explore the linguistic features in image data. Compared to many other studies in linguistic landscape research, which usually consider smaller data sets, our investigation draws on a large corpus of geocoded photos (N = 25,504) and takes an inductive, corpus-linguistics approach. A corpus linguistics approach with its focus on relative frequencies can help us to find out what is recurrent and what is rare in the textification of public spaces. Determining and quantifying linguistic forms and patterns on the basis of a large corpus can uncover linguistic phenomena which would otherwise go unnoticed in analyses of smaller corpora or selected signs (as has prevailed so far in linguistic landscape research). The advantage of quantitative research of this kind is that it gives a more reliable and detailed insight into the "linguistic colonialization of public space" ("Kolonialisierung des öffentlichen Raums durch Schrift" Auer 2010: 295, our translation). Accordingly, text identification software could be a good tool to strengthen linguistic approaches to the analysis of signs in place.

Admittedly, like many corpus-linguistics studies, our analysis is limited as it ignores the non-linguistic, i.e. the spatial, temporal, social and cultural, context of writing on public display and thus cannot explain why a linguistic landscape appears the way it does. But it can be a crucial first step in widening our understanding of 'linguistic' landscaping and may serve as a basis for further investigations combining quantitative methods and qualitative, context-sensitive methods for a deeper understanding of the make-up of specific linguistic landscapes.

This chapter begins with a description of the corpus and data drawn on (section 2). The functions of the software tool Google Cloud Vision API are then introduced and the performance and limitation of its use demonstrated (section 3). In the last section, an evaluation of the usefulness of the Google Cloud Vision API is presented.

2. Corpus and data

The data drawn on to explore the usefulness of automatic label and text extraction with Google Cloud Vision API were collected in the interdisciplinary research project *Metropolenzeichen: Visuelle Mehrsprachigkeit in der Metropole Ruhr / Signs of the Metropolises: Visual multilingualism in the Ruhr Metropolis* (project monograph: Ziegler et al. 2018), a cooperative project that brought together linguists, urban sociologists and integration researchers with a sociopsychological focus. It was based on what is the largest systematic data collection of signage in public space to date, consisting of more than 25,000 image data, 180 narrative interviews with pedestrians and shop owners, and 1,000 telephone interviews. The aim of this project was to investigate the presence, perception and production of visible multilingualism in the public space of the Ruhr Metropolis, an urban agglomeration in the northwest of Germany with approximately 5.3 million inhabitants. As one of Germany's major migrant regions, its history of immigration goes back to the nineteenth century. The development of coal mining and the steel and iron industries increased the demand for workers, which resulted in several waves of labour immigration (cf. Friedrichs 1996: 133–172) and particular settlement patterns.

These settlement patterns[1] have led to residential segregation, "meaning the degree of unequal distribution of the resident population over the territory of a city in terms of social status characteristics (social status of residential areas), of family forms and life styles (family status), and in terms of the ratio of Germans to immigrants" (Strohmeier & Bader 2004[2]). A key feature of residential segregation in the Ruhr Metropolis is the north-south divide along the A 40 motorway, the so-called 'social equator', which divides the cities into ethnically diverse and less diverse, poor and less poor, educated and less educated neighbourhoods (Kersting et al. 2009). The research design of the *Metropolenzeichen* project accounts for this divide by undertaking a cross-sectional study in the cities of Duisburg, Essen, Bochum, and Dortmund. In order to collect comparable data, two neighbourhoods were chosen in each city along the A 40 motorway: Dortmund-Nordstadt, Bochum-Hamme, Essen-Altendorf and Duisburg-Marxloh north of the A 40 motorway; Dortmund-Hörde, Bochum-Langendreer, Essen-Rüttenscheid and Duisburg-Innenstadt south of the A 40 motorway. The following infrastructural units were included: main station, citizens' office one day-care centre per neighbourhood and one cultural institution per city.

These precisely defined areas were fully documented; each individual sign[3] (N = 23,195) in the public space was photographed according to the maxim 1 item, 1 photograph. In addition, in each of the above-mentioned cities the linguistic and semiotic landscapes (outdoor only) were photographed in full at a central station, a tourist attraction, a citizens' office and a children's day-care centre (N = 2,309). This large corpus of 25,504 geocoded digital photographs, taken from September 2012 to December 2013, forms the basis of the *Metropolenzeichen Database*. The photographs stored in the online database are linked to a map of the Ruhr Metropolis to show the distribution and density of visual multilingualism. The image database also provides metadata for the following categories according to Scollon/Scollon (2003) and Backhaus (2007): choice of language/variety, type of discourse (e.g. commercial, regulatory, transgressive, commemorative), type of name (e.g. institution, shop, gastronomy, toponym), information management (e.g. complete, partial, extended translation), appearance, typography (e.g. antiqua, grotesque, black letter, handwriting), and size. This tagging system allows complex search strategies to analyze sociolinguistic and geographical aspects of visual multilingualism.

Identifying languages is central to the tagging of the image data. To assign languages, a distinction was made between "text" (e.g. instruction, warning, advertising...) and "name" in order to take into account the proportion of names and their significance as indicators of visible multilingualism. Accordingly, all names on signs, stickers, posters, etc. were tagged separately. In total, 11,702 (46 %) of the 25,504 geocoded image data in the *Metropolenzeichen* corpus contain proper names.

For the parts defined as "text" there are a total of 27,265 different language occurrences. Around 250 image data contain pieces of text that could not be clearly assigned to a language, mostly proper names, ad hoc word creations or abbreviations. Another 9,952 items (39.5 %) are transgressive items such as stickers, tags and graffiti, which are often designed manually and

[1] For more information on the historical development of settlement patterns in the Ruhr Metropolis cf. Ziegler et al. 2018: 16–53.

[2] Online document, retrieved 4.12.2019 from https://difu.de/publikationen/demographic-decline-segregation-and-social-urban-renewal-in.html.

[3] A "sign" can be defined as any public display of words, characters, numbers, images or symbols to communicate information or attract attention (cf. Shohamy/Gorter 2009: 1). For the purpose of this study, no signs on moving objects such as cars and buses or mobile signs such as T-shirts and bags were included in the corpus.

artistically and are therefore not always easy to decipher. Many texts on signs also show that the notion that one can always clearly distinguish between individual languages does not do justice to the complexity of language use in many cases. Often, for example in language/word plays in the commercial discourse (but not only there), it is not easy to determine whether a text can be assigned to language A or language B, since it has forms of both languages. This kind of written language mixing might result, for instance, from the emulation of communication practices that are typical for bi- and multilingual speakers, such as code switching, which refers to the alternation between two languages within a conversation. These cases were generally taken into account in the analysis. For this reason, the advertising slogan of a Turkish telecommunications company: "Rede mit wem du willst, hem de doya doya."/ lit. "Speak with whom you want, as much as you want" was considered bilingual and correspondingly tagged with the languages German and Turkish. One-word insertions and transfers, such as the use of Polish "Urlop"/"holidays" in the advertising slogan "Schmeckt wie Urlop in der Heimat"/"Tastes like Urlop in the homeland"/"Tastes like holidays in the homeland", were treated in the same way, i.e. tagged as Polish in this case. Another case are linguistic features that are typical of colloquial language use. These features (N = 110) are mostly forms of regional varieties, such as Ruhr German "hömma"/"listen", which were tagged as "non-standard".
The main focus of the present chapter lies on the creation and analysis of the actual text corpus represented in the signs. The functionalities of the Google Cloud Vision API were therefore applied to a sub-corpus of 14,000 geolocalized photos (excluding graffiti).

3. Corpus linguistic exploration

3.1 Functions of the Google Cloud Vision API

Corpus linguistic exploration requires transcription of the textual content of the signs. While this could easily be done manually for a small corpus, it is very time-consuming for several thousand signs. In such cases, automatic processing facilitates corpus creation tremendously. For the present study, the online service 'Cloud Vision API', offered by Google, has been used (https://cloud.google.com/vision/). Based on training data originating from a huge dataset, Cloud Vision API can extract various types of information from image data. Next to the extraction of text, comparable to OCR in the classical scanning process, the service can also provide information about the objects in the image, about landmarks or faces and also about the colors used in the image. In combination, the information extracted automatically out of the image provides a structured way to analyze large image databases. Access to the Cloud Vision API is possible with a range of programming languages (Python, R, Node.js, php, etc.). A certain number of images can be processed free of charge by Google. If this contingent is exceeded, a usage-based fee applies. Note that all images are transferred to a Google server where the text extraction takes place; privacy aspects, e.g. under the 'General Data Protection Regulation' (GDPR), consideration may, therefor, need to be given to privacy aspects.
For the present study, the textual information has been extracted for the sub-corpus of 14,000 signs. Since all analyses were performed using the data science platform 'R' (R Core Team 2019), the R package 'googleCloudVisionR' (Koncz 2019 et al.) was used for text extraction. Developing the processing pipeline in R allowed us to keep text extraction, corpus building and corpus querying in system.
The following four examples (Figure 1 to 4) illustrate the performance of the automatic text extraction for different types of signs. The left-hand column contains the original image and the (green) bounding boxes added automatically by the Cloud Vision software to identify the areas containing text. The right-hand column displays the recognized text for each bounding box.

Figure 1: Example of text extraction result for a German sign

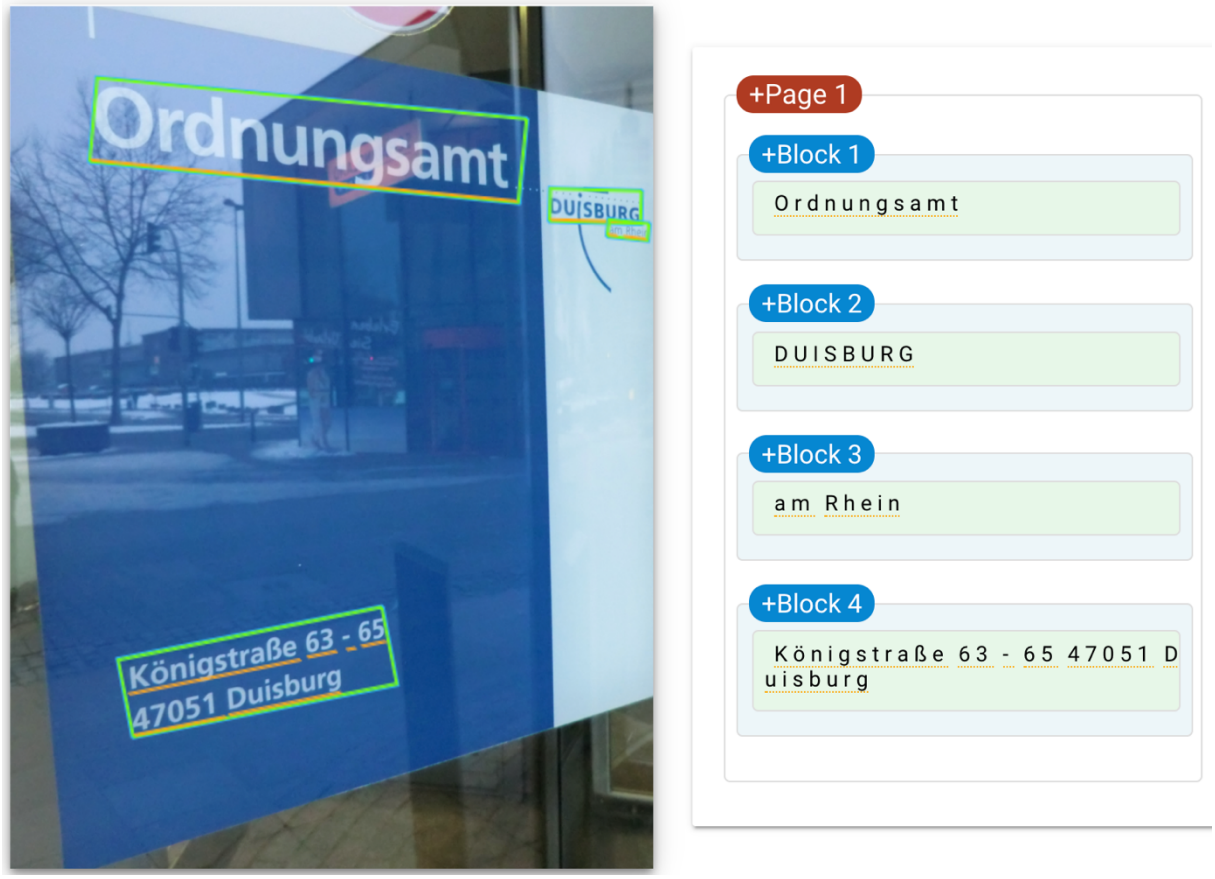


Figure 2: Example of text extraction result for a Turkish sign

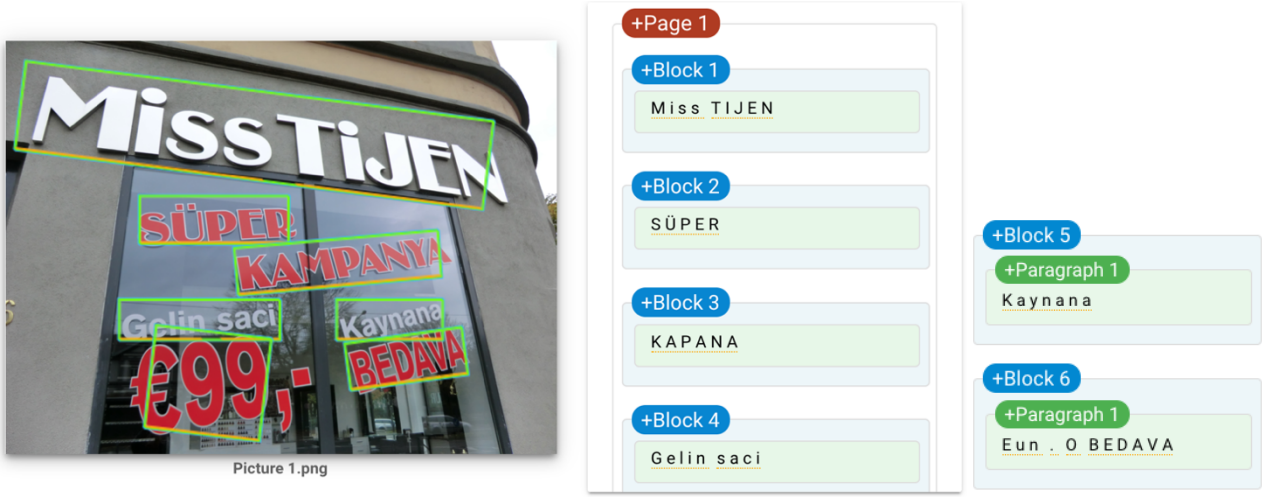


Figure 3: Example of text extraction result for a hand-written sign

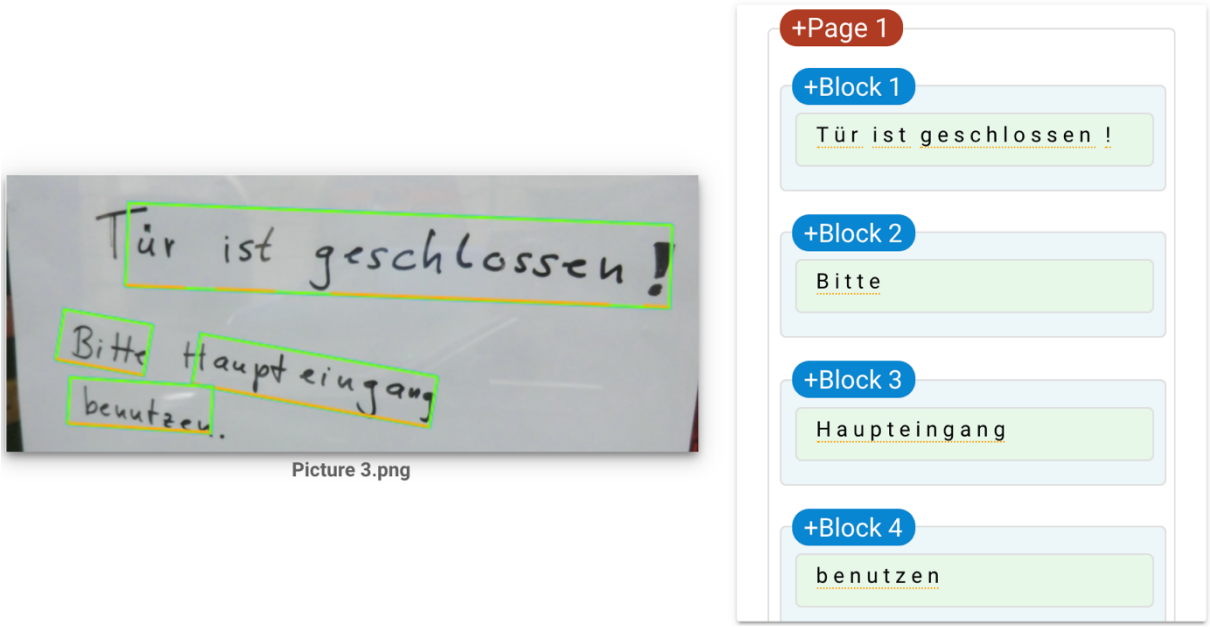


Figure 4: Example of text extraction result for a complex sign

As can be seen, the recognition rate is generally rather high. Due to the integration of language-specific resources and machine learning algorithms, most of the text in the signs could be extracted with relatively high accuracy. Surprisingly, also the hand-written sign (Figure 3) is

recognized without any errors. Only a few problems could be identified: In Figure 1 a line break is not recognized correctly. In Figure 2 the Turkish word 'Kampanya' is recognized falsely as 'Kapana', while 'Bedava' is recognized correctly but the price tag to the left is falsely connected with it. In Figure 4, however, the limitation of the text extraction becomes clear. Due to the curved structure of the sign, the dirt and the additional sticker on top of the original sign, the text extraction performs poorly. Not only is it quite impossible to recognize certain words, the distorted image itself leads to disruption of the lines of text and ultimately to unintelligible word fragments.

Overall, and given the extreme heterogeneity of the image material, it can be concluded that the recognition rate is fairly high. Limitations of Cloud Vision emerge when it comes to bad image quality leading to distorted letters, overlapping graphical elements within the text, varying text size and text direction and usage of non-Latin or artistic and stylized scripts. These difficulties then can lead to well-known OCR errors like the confusion of <l> and <i>, <P> and <R>, <A> and <R> or <z> and <r> (e.g. 'Saiz' instead of 'Salz', 'Rils' instead of 'Pils', 'Rlt' instead of 'Alt', 'Zusatrstoff' instead of 'Zusatzstoff') or to a missing or mis-interpreted umlaut (e.g. 'Konig' instead of 'König', 'w0rzigem' instead of 'würzigem'). Due to its architecture, Cloud Vision is continuously improving its quality by 'learning' from all the images that have been previously analyzed. For the present exploratory study all these errors remain uncorrected, as it was our aim to evaluate the method for a large data corpus even if the data still contain errors.

After applying the Cloud Vision algorithm to the 14,000 images, it was possible to extract a corpus containing approximately 600,000 text tokens. Before utilizing this text corpus for the subsequent linguistic analysis, some general computer-linguistic preprocessing tasks must be carried out, for which the NLP package 'spaCy' with the language model for German was used.[4] In a first step, all punctuation signs, numbers and number-like tokens (e.g. in time specifications) were excluded for this study, as we are mainly interested in the vocabulary used in signs. The corpus is thus reduced to 550,000 tokens, of which the majority, 429,000 tokens, come from the German set of signs. This German subset now forms the basis of further text processing. In the next step, all text tokens were automatically tagged with the corresponding part-of-speech tags[5] (i.e. word classes) and also the lemmatization of the inflected word forms (nouns, verbs, adjectives, pronouns, determiners) was carried out (e.g. tokens *Preises, Preis, Preise* > lemma *Preis*).

These preprocessing steps also allow for a first quantitative evaluation of the recognition quality of the words in the text corpus. To achieve this, we automatically checked the spelling of all words by applying the German dictionary of the 'Hunspell' spellchecker.[6] In the overall result, 71 % of the 429,000 were tagged as correct words of German. The rate of correctness varies considerably across word classes: the closed word classes of determiners, pronouns, conjunctions, subjunctions, particles and auxiliaries have rates above 95 %, while nouns, verbs and adjectives range between 53 % and 71 %. A closer look at the nouns classified as false

[4] 'spaCy' is an industrial-strength standard Python package for all kinds of natural language processing tasks and available via https://spacy.io/. All NLP tasks for this study have been conducted, however, on the R platform, using 'spacyR', 'tidytext' and the 'tidyverse' packages.

[5] The set of part-of-speech tags originates from the Universal Dependencies initiative (https://universaldependencies.org/u/pos). The algorithm for part-of-speech tagging is based on a large morphologically annotated dictionary of German. Due to a missing syntactic component one should be aware that errors might occur, e.g. for capitalized verbforms or multi-word expressions.

[6] 'Hunspell' is an Open Source software package for spellchecking available via http://hunspell.github.io/.

reveals that most of them are indeed correct words of German but are simply missing from the spelling dictionary because most of them are local toponyms (names of cities, villages, streets, districts), brand names (from shop signs), personal names (of medical doctors, lawyers) or local abbreviations (VRR, EWAG, WAZ, S2, etc.). Adding these words to the correct ones would clearly increase the rate of correctness even further. It thus seems safe to assume that the correctness of text extraction by Google Cloud Vision for this image corpus is somewhere above 75 %. Based on these results, the corpus linguistic analysis can now take place.
Table 1 gives an overview of the word classes in the annotated corpus and the corresponding frequencies for tokens. As one would expect, nouns represent the largest share of tokens on the signs, and the frequencies of the remaining word classes are considerably lower.[7] Note that the label 'unknown' contains words which could not be automatically assigned to a word class.

Table 1: Overview of the word classes and the number of word tokens in the text corpus

NOUN	267428
ADJ	31505
ADP	34178
unknown	14952
DET	22425
VERB	20395
CONJ	10606
ADV	10100
PRON	8667
AUX	4573
PART	3011
SCONJ	855
	428,695

3.2 Corpus linguistic case studies

In order to obtain a corpus-linguistic overview of the grammatical and lexical structure of the signs, the following analysis focuses on the most frequent words per word class, the most relevant words per sub-corpus, the average word count and character count per sign, and finally on some frequent trigrams.

Most frequent content words

The first analysis is intended to give insight into the most frequently used content words in the text corpus. Due to the part-of-speech tagging it is possible to easily generate frequency lists for the various word classes. Table 2 lists the 30 most frequent nouns, presenting an instant insight into the most prominent thematic areas on public signs.

Table 2: The 30 most frequent nouns in the text corpus

Uhr	3781	Essen	1288	Dortmund	871
Telefon	1400	Duisburg	881	Bochum	838

[7] For a manual analysis of word classes in German signage in Essen-Altendorf showing similar results to those described above cf. Schmitz (in press).

Jahr	778	Donnerstag	367	Sonntag	319
Straße	727	Information	366	Stunde	305
Samstag	648	Service	362	Vereinbarung	298
Montag	608	Monat	347	AG	296
Euro	582	Stadt	336	Ruhr	296
GmbH	550	Salat	328	Mittwoch	293
Freitag	472	Tag	324	Bank	290
Kind	405	Preis	323	Zeit	260

Starting with the most common noun, *Uhr* 'clock', it is apparent that time specifications form a first cluster of topics on the signs (cf. also *Jahr* 'year', *Samstag* 'Saturday', *Montag* 'Monday', *Freitag* 'Friday, *Donnerstag* 'Thursday', *Monat* 'month', *Tag* 'day', *Sonntag* 'Sunday', *Stunde* 'hour', *Mittwoch* 'Wednesday' and maybe also *Vereinbarung* 'arrangement'). Next, names for localities and other place names occur frequently (toponyms like *Essen, Duisburg, Bochum, Ruhr*; *Straße* 'street', *Stadt* 'city'). Finally, nouns such as *Euro, GmbH* 'LLC', *Salat* 'salad', *AG* 'corporation' and *Bank* suggest the sphere of commerce and restaurants.[8]

As for the verbs (Table 3), lemmatized to the infinitive, one can observe typical actions in public space such as *schließen* 'to close', *freihalten* 'to keep clear', *verbieten* 'to prohibit', *öffnen* 'to open', *fahren* 'to drive', *beachten* 'to respect, *einwerfen* 'to throw sth. in' as in "Keine Werbung einwerfen"/"No advertising material", or *abschleppen* 'to tow away'.

Table 3: Most frequent verbs

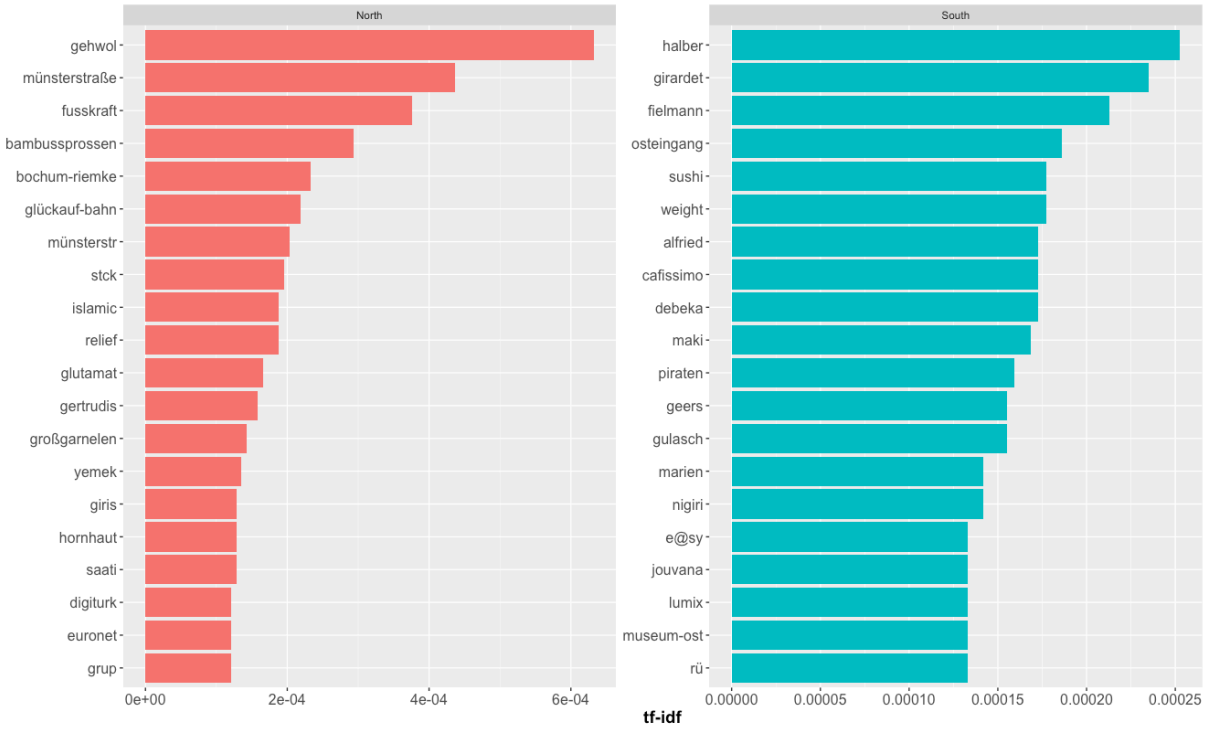
können	665	freihalten	129	betragen	101
machen	246	gehen	127	erreichen	98
gelten	234	verbieten	121	sollen	98
finden	221	öffnen	118	einwerfen	94
erhalten	205	suchen	117	nutzen	91
geben	168	fahren	116	halten	90
wollen	156	entnehmen	110	abschleppen	89
stehen	148	beachten	109	sichern	83
schließen	145	freuen	109	benutzen	82
kommen	132	bieten	106	statten	82

While these frequency lists offer a first insight into the most common semantic and linguistic concepts used on signs in the public space, we propose a more sophisticated analysis of the text corpus by also taking the geographical space into consideration. Following one of the main hypotheses of the *Metropolenzeichen* project, the Ruhr Metropolis is characterized by a north-south divide, which is related to linguistic and cultural differences as the diversity of languages, in particular migrant languages, is higher in the districts north of the A 40 motorway (high ethnic diversity) than in the districts to its south (low ethnic diversity). In order to explore this divide, the text corpus is split up according to these geographic regions and expanded to also include all non-German signs. A commonly used statistical weight for text mining is then applied, i.e. the 'term frequency-inverse document frequency' weight (tf-idf), to discover the

[8] The Turkish noun *taksit* 'instalment', which was often used on advertisements and for Turkish banks, is the only non-German word among the most frequent words.

most relevant words per sub-corpus. According to this weight, the statistical influence of extremely frequent words is decreased and that of rather special words for a sub-corpus is increased, subsequently revealing prevalent semantic concepts for the sub-corpora. Figure 5 lists the 20 words per region with the highest scores for tf-idf, providing evidence that the two regions are indeed distinguished by different semantic concepts. As one would expect, for both regions characteristic toponyms are high among these most relevant words, e.g. *Münsterstraße, Bochum-Riemke, Glückauf-Bahn* for the north, or *Girardet, Alfried (Krupp), Museum-Ost, Rü(ttenscheid)* for the south. On the other hand, the remaining words do not immediately indicate a link to the region.
Strikingly, in the north, four Turkish words can be found, i.e. *yemek* 'food', *giris* 'entrance', *saati* 'time' and *grup* 'group'; this correlates with the high native Turkish population in this area, which has a share of 44 % in Duisburg-Marxloh and 28.9 % in Dortmund-Nordstadt (neighbourhoods located north of the A 40) compared to only 7.8 % in Essen-Rüttenscheid and 13.3 % in Duisburg-Innenstadt (neighbourhoods located south of the A 40). Consequently, no Turkish words are found for the south.

Figure 5: Term frequency-inverse document frequency weight (tf-idf) for the investigated neighbourhoods north and south of the A 40 motorway



However, caution is indicated when explaining these frequency patterns, as a possible bias originating from varying numbers of images per location cannot be ruled out. The brand names *Gehwol* and *Fusskraft*, for example, have gained considerable influence for the north because of frequent occurrences on several images for the same shop. The same holds true for *islamic* and *relief*, which are found on copies of a newsmagazine article on display on many store windows in the northern region.

Average word count and character count per sign

The next feature addresses the mean length of text on a sign. As for the averages for word count and character count, some characteristic differences emerge between the two locations. The box plots in Figure 6 and Figure 7 reveal a broad variation of text length for all locations, thus

creating considerable overlap. However, for both word count and character count, the mean (black vertical line) is highest for signs from the south, where one thus finds the longest texts on signs.

Figure 6: Average word count per sign and per region

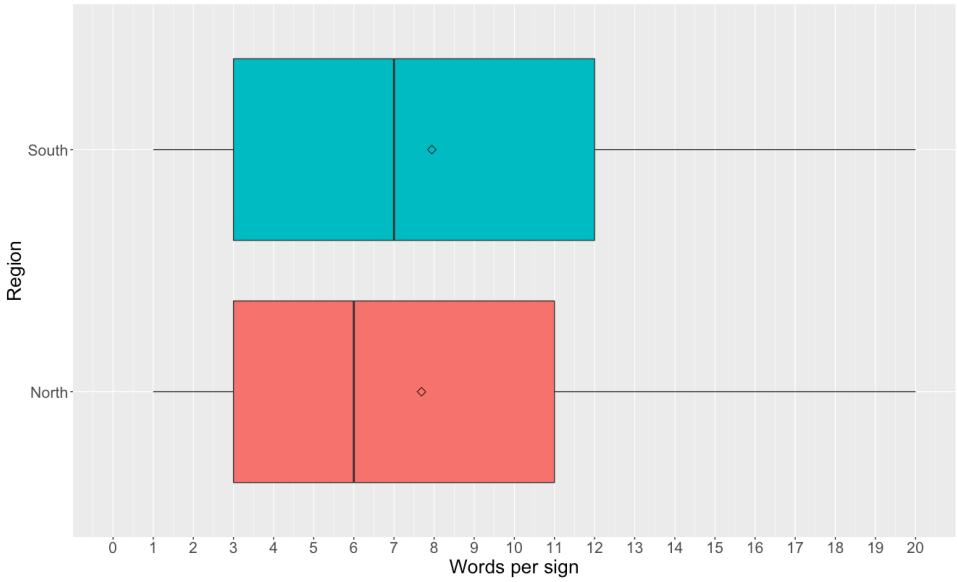
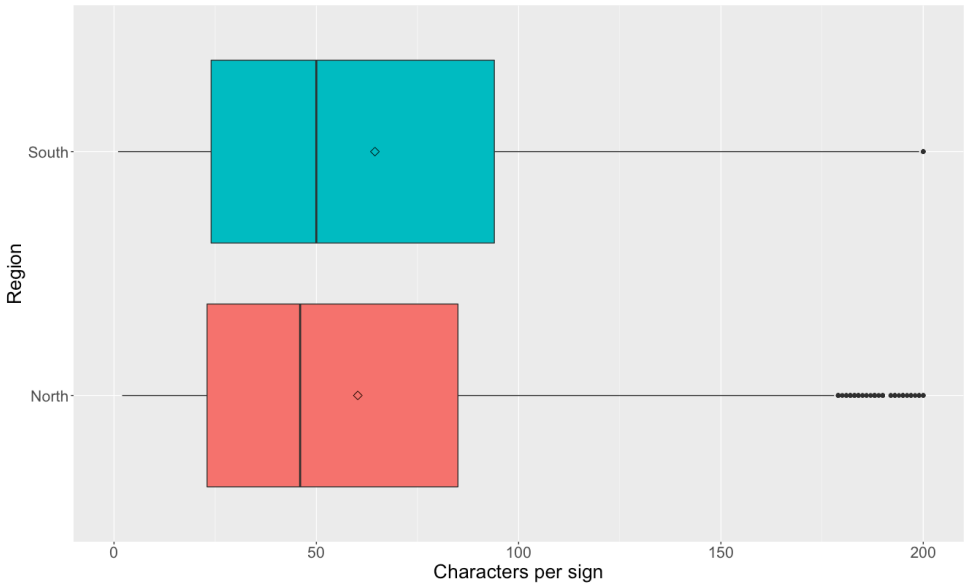


Figure 7: Average character count per sign and per region



Frequent N-grams

In order to gain an insight into the preferred word sequences (so-called N-grams), a 4-gram analysis was carried out. The analysis reveals (cf. Table 5) that the 4-gram *Bank an ihrer Seite* (43) 'bank at your side' is the most common, followed by the warning *Fahrzeuge werden kostenpflichtig abgeschleppt* (40) 'vehicles will be towed away at the owner's expense', the indication of opening hours *von x:xx uhr bis x:xx uhr* (39) 'from x:xx to x:xx', and the thank-you note *vielen Dank für Ihr* (34) 'many thanks for your'. The most common 4-grams are thus to be found in the commercial and regulatory discourse, i.e. in contexts involving advertising and, in the broadest sense, instructions.

Table 5: Most frequent 4-grams

- bank an ihrer seite (43)
- fahrzeuge werden kostenpflichtig abgeschleppt (40)
- von x:xx uhr bis x:xx uhr (39)
- vielen dank für ihr (34)
- dienstag mittwoch donnerstag freitag (28)
- wir freuen uns auf (28)
- bahnhof non smoking station (27)
- termine nach vereinbarung tel (27)
- bewohner mit parkausweis museum (25)
- rauchfreier bahnhof non smoking (25)
- nutzung ist rechtswidrig und (24)
- lesbar hinter die windschutzscheibe (23)
- dank für ihr verständnis (21)

3.3 Geographical mapping of linguistic units

In order to find out whether the distribution of specific words is sensitive to the spatial and social characteristics of the Ruhr Metropolis, i.e. to the already mentioned north-south divide, we were interested in how certain keywords from the regulatory domain were distributed in space. The analysis thus focuses on the regional patterns for the lexemes *verboten* 'prohibited, *Polizei* 'police' and *beachten* 'to respect'. Due to the geocoding of every image, it was then easy to map signs with the respective keywords to the geographical space. The results in Figure 8 show a clear pattern: by far the most occurrences, i.e. 37 out of a total of 216, can be found in the north of the city of Dortmund (Dortmund-Nordstadt). It is also noticeable that the north-south difference in Dortmund is significantly greater than in the cities of Bochum, Essen and Duisburg. For example, there are only 5 occurrences for prohibition lexemes in the southern neighbourhood of Dortmund (Dortmund-Hörde). It seems that the need to regulate actions in the public space is higher in Dortmund-Nordstadt than in Dortmund-Hörde and the other neighbourhoods investigated. According to Mpendukana (2014: 464), linguistic landscape is a "space made meaningful as a result of human interactions, inscriptions, buildings and other structures; it encompasses the accumulation of social practices and historical meanings which people attach to that space". In this perspective, a higher frequency of prohibition signs indicates an increased need of state authorities, private companies or private individuals for protective measures against others' claims to use the space. In that sense, the larger number of prohibition signs semiotically differentiates Dortmund-Nordstadt from other neighbourhoods in Dortmund and beyond.

Figure 8: Geographical distribution of *verboten* 'prohibited', *Polizei* 'police', *beachten* 'to respect'

4. Evaluation of application usability

In our study, we were able to show that the application of automatic text extraction of a large corpus of public signs opens the door to corpus linguistic analysis in linguistic landscape research. The corpus-linguistic case studies carried out for the most frequent content words and frequent N-grams showed that time specifications figure most prominently in the data, along with commerce- related words and word combinations. Moreover, the analyses for the average word count and character count per sign revealed differences that would not have been as easy to detect manually and that relate to the north-south divide in the Ruhr Metropolis. With regard to text length, there is a tendency to address potential recipients in the neighbourhoods south of the A 40 with longer texts. Whether this is due to linguistic strategies of politeness or different fields of advertising should be the subject of further investigation. With regard to discourse types, no differences were found in the infrastructural discourse due to the fact that these texts are generally short and highly standardized. Rather, differences occurred where text producers are free to choose their linguistic strategies, as is the case in the commercial discourse.

Finally, we demonstrated how automatic text extraction can be combined with geosemiotics by analyzing selected keywords geographically and visualizing their distribution on a map. So far, geovisualization has been carried out for languages and language clusters only, not for linguistic elements, which adds a micro-linguistic perspective to the spatial distribution of linguistic phenomena.

The benefit of text extraction analysis is that it opens linguistic landscape research for corpus linguistics by providing a quantitative bottom-up approach.[9] The analysis of large collections of data will yield new research interests such as: (i) occurrence of punctuation marks, constructions, writing systems, genres; (ii) dominance of forms, colours, landmarks, faces; and (iii) comparison of linguistic structures and languages.

Acknowledgements

We acknowledge financial support from 1 August 2013 to 31 August 2018 from the Mercator Research Center Ruhr (MERCUR) for our project *Metropolenzeichen: Visuelle Mehrsprachigkeit in der Metropole Ruhr*/*Signs of the Metropolises: Visual multilingualism in the Ruhr Metropolis* (PI: Evelyn Ziegler; MERCUR reference number: Pr-2012-0045).

5. References

Androutsopoulos, Jannis (2014). "Computer-mediated communication and linguistic landscapes". In Janet Holmes & Kirk Hazen. (eds.), *Research methods in sociolinguistics: A practical guide*. Wiley-Blackwell: Oxford, 74–90.

Auer, Peter (2009). "Visible dialect". In H. Hovmark, I. Stampe Sletten & A. Gudiksen (eds.), *I mund og bog. 25 artikler om sprog tilegnet Inge Lise Pedersen på 70-årsdagen d. 5. juni 2009*. Kopenhagen: Nordisk Forskningsinstitut, 31–46.

Auer, Peter (2010). "Sprachliche Landschaften. Die Strukturierung des öffentlichen Raums durch die geschriebene Sprache". In Arnulf Deppermann & Angelika Linke (eds.), *Sprache*

[9] The described text extraction techniques will further broaden the horizon of Linguistic Landscape research. As a next step, it is envisaged to include the Google Cloud Vision API into the 'Lingscape' application, which was developed at the University of Luxembourg (cf. Purschke 2018) to compile a multilingual text corpus.

intermedial – Stimme und Schrift, Bild und Ton. de Gruyter: Berlin (= IDS-Jahrbuch 2009), 271–300.

Backhaus, Peter (2007). *Linguistic Landscapes: A Comparative Study of Urban Multilingualism in Tokyo*. Multilingual Matters: Clevedon.

Blommaert, Jan (2013). *Ethnography, superdiversity, and linguistic landscapes: Chronicles of complexity*. Multilingual Matters: Bristol.

Buchstaller, Isabelle & Seraphin Alvanides (2017). "Mapping the linguistic landscapes of the Marshall Islands". *Journal of Linguistic Geography* 5 (2): 67–85.

Domke, Christine (2014). *Die Betextung des öffentlichen Raumes. Eine Studie zur Spezifik von Meso-Kommunikation am Beispiel von Bahnhöfen, Innenstädten und Flughäfen*. Winter: Heidelberg.

Friedrichs, Jürgen (1996). "Intra-regional polarization: Cities in the Ruhr Area". In: John O'Loughlin & Jürgen Friedrichs (eds.), *Social Polarization in Post-Industrial Metropolises*.: de Gruyter: Berlin, New York, 133–172.

Gaiser, Leonie & Yaron Matras (2016). *The spatial construction of civic identities: A study of Manchester's linguistic landscapes*. http://mlm.humanities.manchester.ac.uk/wp-content/uploads/2016/12/ManchesterLinguisticLandscapes.pdf**.**

Garvin, Rebecca T. (2010). "Responses to the linguistic landscape in Memphis, Tennessee: An urban space in transition". In Elana Shohamy, Eliezer Ben-Rafael & Monica Barni (eds.), *Linguistic Landscape in the City*. Multilingual Matters: Bristol, 252–271.

Gilles, Peter & Evelyn Ziegler (2019). "Linguistic Landscape-Forschung in sprachhistorischer Perspektive: öffentliche Bekanntmachungen in der Stadt Luxemburg im langen 19. Jahrhundert". *Zeitschrift für germanistische Linguistik* 47/2: 385–407.

Hennig, Mathilde (2010). "Grammatik multicodal: Ein Vorschlag am Beispiel ortsgebundener Schriftlichkeit". *Kodikas/Code. Ars Semeiotica*: 73–88.

Kersting, V., C. Meyer, K.P. Strohmeier & T. Teerporten (2009). "Die A 40 – Der ‚Sozialäquator' des Ruhrgebiets". In: Prossek, A., H. Schneider, H.A. Wessel, B. Wetterau & D. Wiktorin (eds.), *Atlas der Metropole Ruhr*. Emons: Köln, 142–145.

Koncz, Tamas, Varkoly, Balazs, Lukacs, Peter & Eszter Kocsis (2019). *googleCloudVisionR: Access to the 'Google Cloud Vision' API for Image Recognition, OCR and Labeling. R package version 0.1.0.9000.* (https://cran.r-project.org/web/packages/googleCloudVisionR/index.html)

Mpendukana, Sibonile (2014). "Linguistic Landscapes". In Zanni Bock & Mheta Gift (eds.), *Language, Society and Communication.* Pretoria: Van Schaik, 463–484.

Peukert, Hagen (2013). "Measuring Language Diversity in Urban Ecosystems". In: Joana Duarte and Ingrid Gogolin (eds.), *Linguistic super-diversity in urban areas: research approaches*, John Benjamins, Amsterdam, 75–95.

Purschke, Christoph (2018). “Sprachliche Vielfalt entdecken mit der Lingscape-App”. *Der Deutschunterricht*: 70–75.

Pütz, Martin & Neele Mundt (2019). “Multilingualism, Multimodality and Methodology: Linguistic Landscape Research in the Context of Assemblages, Ideologies and (In)visibility: An Introduction”. In Martin Pütz & Neele Mundt (eds.), *Expanding the linguistic landscape: Linguistic diversity, multimodality and the use of space as a semiotic resource*. Multilingual Matters: Bristol, 1–22.

R Core Team (2019). R: *A language and environment for statistical computing. R Foundation for Statistical Computing*, Vienna, Austria. (https://www.R-project.org/).

Reh, Mechthild (2004). “Multilingual writing: A reader-orientated typology – with examples from Lira Municipality (Uganda)”. *International Journal of the Sociology of Language 170:* 1–41.

Scollon, Ron & Suzie Wong Scollon (2003). *Discourses in Place: Language in the Material World*. Routledge: London.

Schmitz, Ulrich (in press). “Wörter an der Ruhr. Zur Banalität von Straßentexten”. In Derya Gür Şeker (ed.), *Wörter, Wörterbücher, Wortschätze und Wortschatzdidaktik*. (Korpus-) Linguistische Perspektiven. UVRR: Duisburg.

Schmitz, Ulrich (2017). “Randgrammatik und Design”. *IDS Sprachreport* H. 3/2017: 8–17.

Sebba, Marc (2013). “Multilingualism in written discourse: An approach to the analysis of multilingual texts”. *International Journal of Bilingualism* 17.1: 97–118.

Shohamy, Elana & Durk Gorter (2009). “Introduction”. In: Elana Shohamy & Durk Gorter (eds.) *Linguistic Landscape. Expanding the Scenery*. Routledge: New York & London, 1–10.

Strohmeier, Klaus Peter & Silvia Bader (2004). “Demographic decline, segregation, and social urban renewal in old industrial metropolitan areas”. *Deutsche Zeitschrift für Kommunalwissenschaften* (https://difu.de/publikationen/demographic-decline-segregation-and-social-urban-renewal-in.html).

Van Mensel, Luk, Mieke Vandenbroucke & Robert Blackwood (2016). “Linguistic Landscapes”. In Ofelia Garcia, Max Spotti & Nelson Flores (eds.), *The Oxford Handbook of Language and Society*. Oxford University Press: Oxford, 423–450.

Ziegler, Evelyn, Heinz Eickmans, Ulrich Schmitz, Haci-Halil Uslucan, David H. Gehne, Sebastian Kurtenbach, Tirza Mühlan-Meyer & Irmi Wachendorff (2018). *Metropolenzeichen: Atlas zur visuellen Mehrsprachigkeit der Metropole Ruhr*. UVRR: Duisburg.