Digitising Collections of Historical Linguistic Data: The Example of The Linguistic Atlas of Scotland

John Monfries Kirk, Christian Hessle

Digitising Collections of Historical Linguistic Data: The Example of The Linguistic Atlas of Scotland

Christian Hessle & John Kirk
University of Vienna

Abstract

The paper provides a preliminary exploration of the possibilities and prequisites for digitising the lexical material of the Linguistic Atlas of Scotland. The survey by written questionnaire on which the atlas is based and also the editing and cartography of the published maps are each introduced and critiqued. Three prototypical maps for the North mid-Scots dialect area are presented. Their lexical content is discussed, especially the issue of lexical categorisation and the representation of extra-linguistic information. The mapping process is then presented, together with a discussion of various decisions which had to be made. The articles ends by recognising two central prerequisites which affect data input: data normalisation and machine readability. In these ways, the paper offers a critical perspective of the digitisation task, ahead of the full national coverage.

keywords

lexis, categorisation, synonymy, interactive maps, interactivity, cartography, linguistic topography, language continuum, visual representation, data normalisation, machine interpretation

1. INTRODUCTION

The age of digitalisation opens up new perspectives on linguistic geography. Thereby, the availability of a broad range of cartographical resources as well as digital visualisation technologies provide a convenient opportunity for a reinterpretation of historical data-sets. Such new opportunities, however, raise questions about the nature of data collections as well as about their topography and cartography in digital environments.

This paper examines perspectives on the digitisation and reinterpretation of historical linguistic data using the example of *The Linguistic Atlas of Scotland* (Mather & Speitel 1975, 1977). After an

introduction to the atlas, covering its contents and design as well as its shortcomings, a prototype project from earlier this year entitled *Towards a Digital Version of The Linguistic Atlas of Scotland* (Hessle 2019) is presented. Thereby, a focus is laid on the lexical analysis of informants' responses, on the prerequisites under which a categorisation of the results can be established, and on how the data can be represented visually. The third part of the present paper shows how historic data collections can be digitally processed and thus touches on the limitations of data normalisation and machine readability.

2. THE LINGUISTIC ATLAS OF SCOTLAND

The first volume of *The Linguistic Atlas of Scotland* by James Y. Mather and Hans-Henning Speitel was published in 1975. The data-set of the survey is based on a questionnaire that was sent out in 1952 (cf. Mather & Speitel 1975: 379) to residents of Scotland, the northern English counties of Cumberland and Northumberland, Northern Ireland and the county Donegal in the Republic of Ireland (ibid. 8). The informants were chosen by local "headmasters of primary schools" who were asked to select "middle aged or older and a lifelong inhabitant[s]" (ibid. 14) with a focus on rural areas. In the questionnaire, the informants were asked for a "word or words commonly used for [Standard English items] in [their] own locality" (ibid. 13). All in all, the first volume of the atlas includes responses by 1,774 informants (Mather & Speitel 1977: 9). The results are presented on 122 linguistic maps and list for 90 lexical items (Mather & Speitel 1975: Contents). Moreover, the volume includes an introduction, a facsimile of a sample questionnaire, 21 phonetic and orthographical maps, a key map of the informants' localities, a list of all informants, a county map, a population density map from 1951 and a physical map of Scotland (ibid.). In 1977, a second volume of the atlas was published, including 80 lexical items and 832 informants (cf. Macaulay 1979: 224-225). Taken together, both volumes of the atlas provides sources for 226,220 responses.

Many will concur with McClure (1976: 233) that "the *Linguistic Atlas of Scotland* is by any standards a monumental work of scholarship and a major contribution, not only to Scottish dialect studies, but to dialect research throughout the English-speaking world and to theoretical dialectology". Despite such general appreciations of the impressive scope of *The Linguistic Atlas of Scotland* (cf. also Macaulay 1977, 1979, 1985, Millar 2018: 123-127, Murison 1978), it is hard to avoid considering some of the linguistic decisions taken by Mather and Speitel in a critical perspective. In this respect, Derrick McClure (1975: 227) emphasises that "no means were provided of determining the correct choice [...] between three possible interpretations of an informant's

failure to respond to an item in the questionnaire". Hence, it remains unclear whether the informant "did not know the dialect word required, no dialect word existed in his [or her] locality, [or] he [or she] failed to understand the question" (ibid.). Moreover, McClure points out that "[v]ery similar orthographic forms are in many cases presented separately" (ibid. 229), while hapax items that are "attested only once" (ibid. 230) are generally not represented on the maps. Many of these hapaxes are simply further orthographic variants of words which are indeed mapped. A study of the East Central Scots responses shows that 483 or 51.2% of the data are indeed unmapped hapaxes (Kirk 1994a: 57), when it would surely have been appropriate to treat them as orthographic synonyms. McClure (1975: 230) also notes that "distinctions of meaning" indicated by the informants are not visualised. Ron Macaulay (1985: 175) insists that "the respondents were asked to supply the local word" and therefore suggests that answers containing the given English word should "be treated as a 'nil' response" (ibid.). In fact, the problem seems to arise from the questionnaire seeking two separate responses: one or more "usual local word(s)" [converted to lower case] (Mather and Speitel 1975: 11), and one or more "less common local word(s)" [converted to lower case] (ibid.). While some of Mather and Speitel's decisions might appear rather questionable, others can be seen as concessions to the physical limitations of a printed atlas.

The visual representation of data in *The Linguistic Atlas of Scotland* has been subject to criticism as well. McClure (1975: 230) criticises the readability of maps for lexical items such as *youngest of a brood* (Map 65) and *splinter* (Map 4), in which "different hatchings are superimposed". Furthermore, Mather and Speitel's methodological approach towards constructing isogloss boundaries remains unclear. On the one hand, the authors describe an isogloss "as a line that surrounds an area in which a particular linguistically defined phenomenon (or sets of phenomena) is found. Outside the isogloss the particular phenomenon is (a) absent or (b) does not form a coherent linguistic area i.e., it is not sufficiently concentrated" (Mather & Speitel 1975: 8). Thus an isogloss may be taken to indicate a "perimeter boundary" (Kirk 1994b: 2368) of the area in which a form occurs, what Kretzschmar (1992: 227) calls "a limit of occurrence". At the same time, Mather and Speitel (1975: 8) claim that isoglosses "often follow geographical contours". When comparing Macaulay's interpretation of isogloss boundaries for Scots dialect items referring to Standard English *splinter* (cf. 1985: 175-180) with the respective lists of responses provided by Mather and Speitel (1975: 158), a discrepancy becomes apparent. As Macaulay (1985: 175) assumes that "a concentration of a particular response [...] [is] clearly outlined", it is not possible for him to identify *spale* as the dominant dialect item for *splinter* in the mountainous parts of Perthshire north of the highland line (cf. 178). Moreover, the interpretation of isogloss lines might lead to the conclusion

that *skelf* can be found in the whole province of Fife (cf. 179), whereas the reference lists show that the item's presence concentrates on the southern coast of Fife, on seaside towns such as Perth and Sterling and to Flandern Moss National Nature Reserve in Perthshire (Mather & Speitel 1975: 158). As a result, conclusions drawn from isogloss or distributional-boundary lines must be treated with great caution.

3. Towards a Digital Version of *The Linguistic Atlas of Scotland*

In January 2019, the unpublished study entitled *Towards a Digital Version of The Linguistic Atlas of Scotland* was completed (cf. Hessle 2019). The investigation includes three of the questionnaire items, namely 'ankle', 'splinter', and 'youngest of a brood' and is restricted to 182 informants from pre-1975 Scottish counties of Clackmannan, Fife, Kinross and Perth, compromising the main distribution area of the North Mid-group of Scots dialects (cf. Johnston 1997: 438). That study's main goal is to outline perspectives on a digitalisation of *The Linguistic Atlas of Scotland* (cf. Hessle 2019: 4) with a focus on reviewing the data-set and its lexical categories. The study combines linguistic methods with digital cartography technologies in order to create individual online maps for three lexical items (cf. Maps 1, 2 and 3, from Hessle 2019). Thereby, each item is mapped twice, allowing the user to choose either between the sex of the informants, or to select age groups. From the cartographical display of identically-coloured circles the topographical extent of any item (or group of lexicalised items) may be inferred without the need for perimeter isoglosses. Hessle's study provides an outlook on how a future digital version of *The Linguistic Atlas of Scotland* can be realised, and which challenges might arise during the process of digitisation.

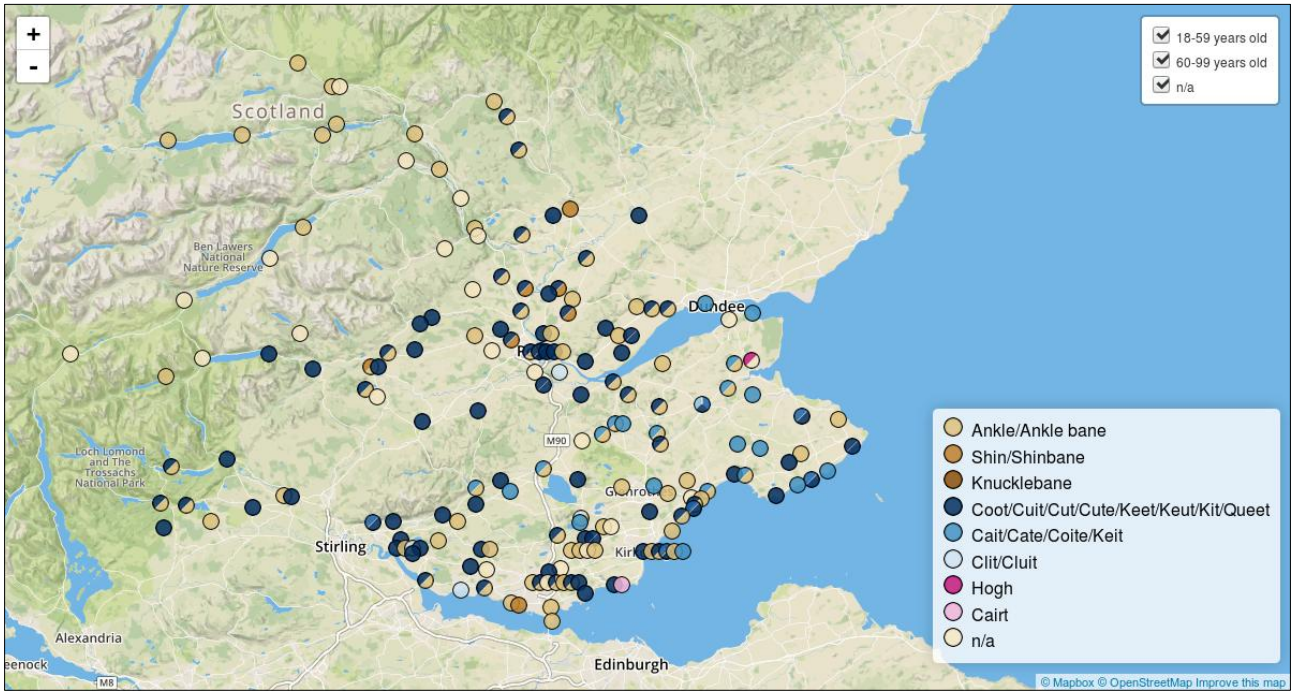

Figure 1: Scots synonyms for SSE 'ankle' (sorted according to age groups) in Clackmannan, Fife, Kinross & Perth (Map 1 from Hessle 2019, accessible online at http://16levels.org/las/ankle_age.html)

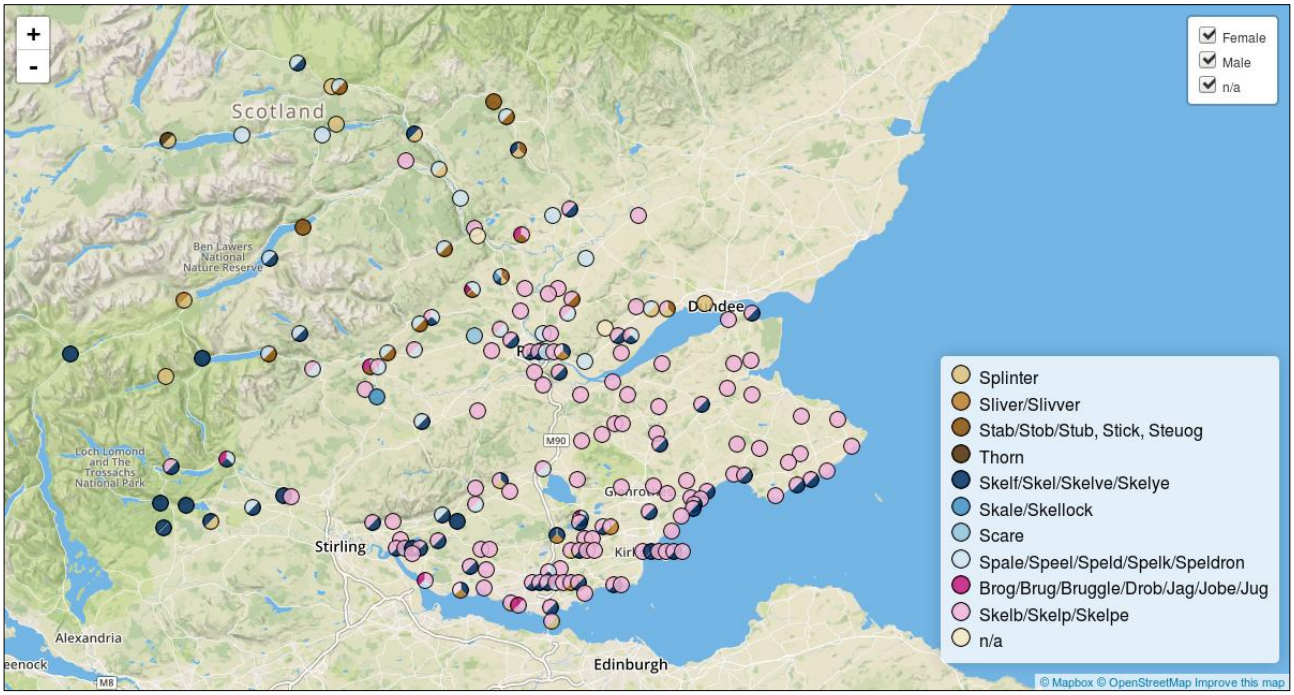

Figure 2: Scots synonyms for SSE 'splinter' (sorted according to gender) in Clackmannan, Fife, Kinross & Perth (Map 2 from Hessle 2019, accessible online at http://16levels.org/las/splinter_gender.html)

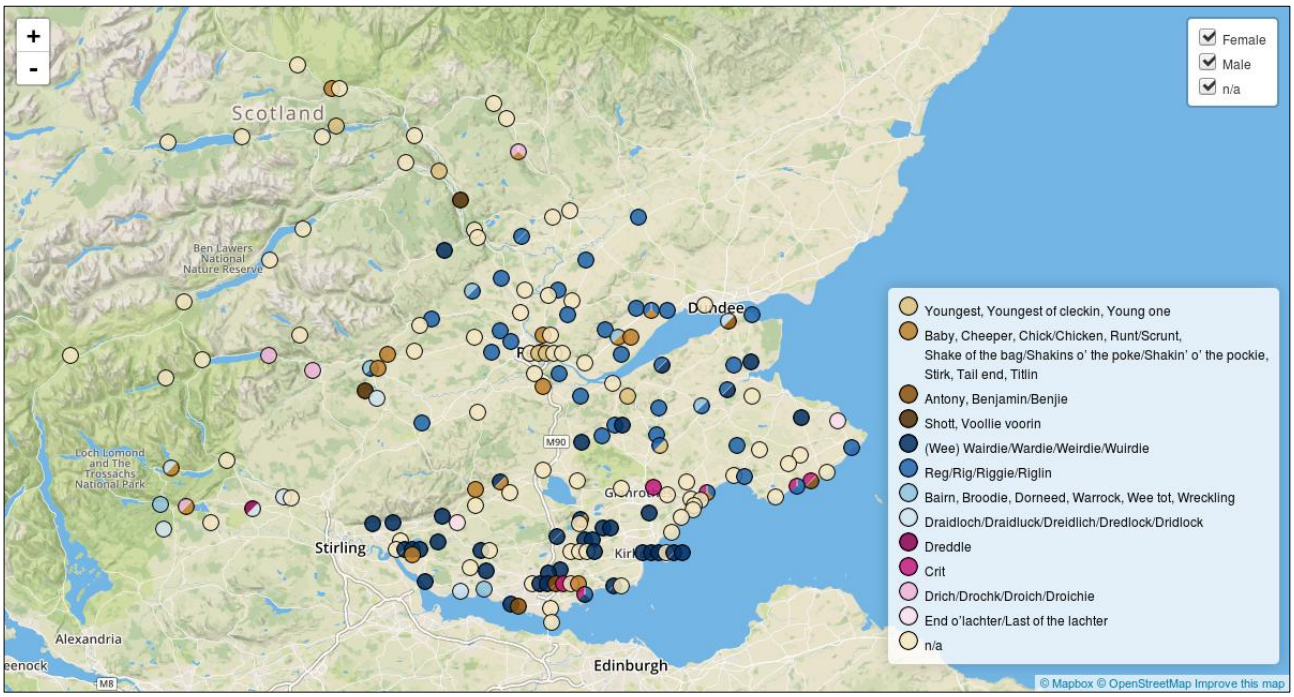

Figure 3: Scots synonyms for SSE 'youngest of a brood' (sorted according to gender) in Clackmannan, Fife, Kinross & Perth (Map 3 from Hessle 2019, accessible online at http://16levels.org/las/youngest_gender.html)

3.1. Lexical analysis

The starting point for the study is a close analysis of the informants' responses. For the etymological part of the analysis, the *Concise Scots Dictionary* (Scottish Language Dictionaries 2017), the online *Dictionary of the Scots Language* (comprising the resources of the *Dictionary of the Older Scottish Tongue* and the *Scottish National Dictionary*) (available at http://www.dsl.ac.uk) and *The Shorter Oxford English Dictionary* are consulted (cf. Hessle 2019: 5). In some cases, historico-cultural sources such as paintings and publications in the fields of architecture and history are also employed in order to reveal further details of the etymological background of an item. For example, Hessle (2019: 8) shows that several items relating to the Standard English item *ankle*, such as *cait, cate, coite* and *keit*, refer to the game of curling. While the connection between *ankle* and *curling* remains unclear in the dictionaries sources, the painting "Hunters in the snow" by Pieter Bruegel the Elder from 1565 reveals that "[curling-]stones were often made from animal bones, particularly the ankle bone of horses" (ibid.). On a different occasion, the Dutch influence on the architecture of coastal areas of Fife in form of "typical crow-stepped gable[s]" (Price 2013) proves essential for establishing a distinction between Gaelic *skelb* and Dutch *skelf*, both Scots synonyms for Standard English *splinter* (Hessle 2019: 13). Last but not least, in order to identify different orthographic forms of the same lexical item, Paul Johnston's word-sets model is used (cf. 1997) as a further basis for categorising the informants' responses.

3.2. Categorisation

In order to make the data of *The Linguistic Atlas of Scotland* digitally accessible, it is necessary to categorise the data on a purely linguistic basis. Ideally, such categories should be both complete and consistent, in order to establish for the informants' responses a stable topological space. According to Graham Flegg (1974: 19), "topology may be thought of as the study of non-metric spatial relationships [and their] continuity." [emphasis removed] Moreover, Alain Badiou (2016: 61-62) describes *topos* as "a category in which can be defined a relation similar to the classical relation of belonging, the famous ∈". Unlike a system adhering to the principle of the excluded middle, he argues that topology follows a rather intuitionistic logic (ibid. 62). It thereby allows the construction of coherent systems in which "it is generally not the case that the negation of negation is equivalent to simple affirmation." (ibid. 60). With reference to the atlas data, Badiou's claim reveals a contradiction in the assumed dichotomy between Standard English items and Scots dialect items as suggested by the questionnaires of *The Linguistic Atlas of Scotland*. Unsurprisingly, the linguistic field between English and Scots is characterised by a multiplicity of corresponding variants. In this respect, Badiou (2007: 19) insists that it is essential to "define the rules of correspondence. Everything concerning these rules depends on the *semantics* of the system, on its *interpretation*." And he concludes: "to speak of the meaning of the system is to speak of its various interpretations." (ibid.) And so to understand the nature of the Atlas data as a collective whole, we need to take into account all of it. Both what the items which are marked and those which are not as well as the manner of mapping are clear interpretations – semantic interpretations - of the data. Thus, instead of focussing on distinctive features, what the analysis of different forms of the same item should come to outline is almost certainly a shared language-continuum. Thereby, as Kirk (1994b: 2363) contends, "the role and function of linguistic maps has plainly shifted from the original demonstration of the distribution of individual linguistic items [...] to the use of geography to explain inherent linguistic variation".

On a different occasion, Badiou draws on the comparison between solving a mathematical problem and playing a game of chess: while a detailed and far-reaching knowledge of opening might provide somebody with a strategic advantage, it is in fact "the path to the solution of a problem [...] that makes you touch a real and has a sort of intrinsic complexity." (2016: 63). Likewise, Sonja Amadae (2015: 74) stresses the importance of "Europe's chess-playing culture", for the Wiener Kreis and the proponents of logical empiricism. This is particularly true for John von Neumann, who is probably best known for his contribution to game theory as well as for his involvement in the

RAND Corporation, a military think tank which played a central role for thermonuclear strategy of the USA in the second half of the 20th century (ibid. 73-76). Von Neumann's rejection of the idea of intuitionistic logic is illustrated by his 'minimax'-concept, in which two chess-players "can select a strategy that will secure a minimum security threshold below which the other player cannot force [their opponent]." (ibid. 75) Obviously, a model such as the 'minimax'-concept is not suited to reflect a linguistic reality. Nevertheless, the contradiction between the empirical approach of the questionnaire-method and the intuitionistic logic of a linguistic topography involves two consequences for a linguistic atlas project. Firstly, one must accept that the process of data categorisation will accompany the researcher throughout the course of the study. It is therefore not possible to draft a complete and consistent set of categories in advance that will then serve as a stable basis for the creation of a linguistic atlas. The visual representation can, however, serve as a valuable tool to further refine and adjust such categories. And secondly, despite "a great temptation to export this concept into general epistemology," (Badiou 2007: 19-20), one must be aware that such a model can only be understood as a set of multiple interpretations of reality. Hence, it allows us to "think the relation between a formal system and its 'natural' exterior." (ibid. 18) As a result, it becomes clear that it will only to a limited extent that linguistic maps will serve as the basis for establishing linguistic generalisations.

3.3. The mapping process: technology, colour-palettes and the display of extra-linguistic data

For mapping the data of *The Linguistic Atlas of Scotland*, the survey *Towards a Digital Version of The Linguistic Atlas of Scotland* (cf. Hessle 2019) combines digital technologies that are easily accessible and well documented. As base-map, the Open Street Maps-project (cf. 2018) is chosen. The open source project founded by Steve Coast was "initially focusing on mapping the United Kingdom" (ibid.), hence, the necessary detail for regions covered by the study is ensured. Since the grid provided for the informants' localities in *The Linguistic Atlas of Scotland* does not seem to correspond to any publicly accessible online source, Google Maps (2019) and the online maps provided by the Ordinance Survey (2019) are used to locate informants in cases when the search on Open Street Maps does not provide the desired results. In order to include visual geographical information, publicly available 'tiles' by Mapbox (cf. 2019) are layered on the maps (cf. Hessle 2019: 7). The data is stored in a GeoJSON-file (cf. 2019) whose graphical output can be accessed in a web-browser by executing a JavaScript-code based on Leaflet (2019), "an open-source JavaScript library for mobile-friendly interactive maps". Compared to a database-solution, the combination of GeoJSON and Leaflet has several advantages. To start with, the maps do not require a database

server and can therefore by run on locally in a web-browser. Furthermore, the technologies used are available under an open source license and include extensive documentation. Even more important and in contrast to a database, a solution based on GeoJSON allows ad hoc adaptations of a map's categories without having to alter the structure of a database. Considering the necessity constantly to reconfigure categories during the process of mapping as described above, flexibility remains the main advantage of the approach combining GeoJSON and Leaflet. On the downside, it must be taken into account that all calculations are executed locally by the web-browser. As a result, large data-sets will significantly reduce the performance of the maps. Moreover, both Leaflet and GeoJSON have technical limitations as far as their configurability is concerned. For example, different categories such as the informants' gender or age groups cannot be toggled in the same map, but must be split to two separate instances (cf. Maps 1, 2 and 3, from Hessle 2019). Furthermore, as a result of GeoJSON's list-character and in contrast to database-structures, logical operations cannot be executed. However, for a geolinguistic prototype study, the combination of GeoJSON and Leaflet is an appropriate solution which can be easily implemented.

Apart from the background technologies used, several visual decisions are taken in order to optimise the readability of the study's maps. As Maps 1, 2 and 3 show, the data is displayed by coloured circles with a diameter of 18 pixels. For data-entries containing between two and six lexical items, the circles are split accordingly, while larger numbers of items are simplified in order to ensure readability (cf. Hessle 2019: 6). The circles use "shades of blue, brown and magenta" (ibid.) in order to guarantee that "[p]eople with red–green colour blindness" (Allred, Schreiner & Smithies) who "account for several per cent of the population" (Leck 1994), are able to interpret the maps. Addressing Macaulay's criticism (1985: 175), the Standard English headwords and 'nil'-items "are differentiated with shades of light brown." (Hessle 2019: 6) However, attempting to background some items by using colourless tones implies problematic side-effects. Several attributes commonly associated with the dichotomy of colourful and colourless are described by Roland Barthes (cf. 2005 [1977-1978]: 49-52). In his analysis of altar paintings, Barthes traces a relation between colourfulness and "festival, riches, upper class" (ibid. 50), while "grisaille, monochrome, 'neutral'" (ibid.) are often associated with "quotidian, social uniformity [and] [...] poverty." (ibid.) Furthermore, Barthes insists that "the Neutral is shown in order to hide the colorful. Here we are in an ideology of 'depth,' of the apparent versus the hidden." (ibid.) Clearly, the application of ideological colour judgements to the field of linguistics is problematic. Even though there might be good reasons for moving certain categories to the background of the viewers' perception, it is necessary to be aware of the semantic implications of such a decision, in particular

when the relation between dominant and subordinate varieties are concerned. Barthes sums up the problem in a concise formula: "[t]he hidden = rich, the apparent = poor." (ibid.) With Barthes' interest in the 'neutral', "[t]he grisaille [...] points to another way of thinking the [...] principle of organization" (ibid. 51). While blue and red – the first colours which might spring to mind when thinking about distinguishing between Scots and Standard English – for Barthes represent "the opposition par excellence," (ibid.) "the monochrome (the Neutral) substitutes for the idea of opposition that of the slight difference, of the onset, of the effort toward difference." (ibid.) As a result, the choice of a colour-palette for a linguistic map is not only a merely technical question, but "becomes a principle of allover organization [...] that in a way skips the paradigm" (ibid.) of the semantics of the visual representation of the linguistic data. Barthes suggests to think of this nuances as a moiré-pattern "whose aspect, perhaps whose meaning, is subtly modified according to the angle of the subject's gaze." (ibid.) In conclusion, rather than emphasising opposition (or binary-opposition) with the choice of contrasting colours, a linguistic map, using a thoughtfully chosen colour-palette allowing nuances, may provide its readers with a rich variety of comparable and equivalent semantic connections.

Compared to a printed linguistic map, its digital counterpart facilitates the display of extralinguistic information. While the maps of the study *Towards a Digital Version of The Linguistic Atlas of Scotland* use a topographical background map, Yuchun Xie *et al.* (2013: 306) suggest that also "data on flora, fauna, and population demographics [could be] [...] made available for real-time mapping to base layers." According to Silviu-Ioan Bejinariu and Florin-Teodor Olariu (2017: 15), such options can "contribute to a much better contextualized analysis of [...] linguistic data". In the case of *The Linguistic Atlas of Scotland*, the content of maps provided separately in the appendix, for example the population density map (cf. Mather & Speitel 1975: Contents), could be directly linked with the linguistic data in a digital version of the map. Moreover, Hessle (2019: 6) shows that extralinguistic information can be embedded as a pop-up window, indicating "the lexical items of the informant's response, a code to identify the informant on the list, their gender and age as well as additional information from the lists [and results from] [...] the research process." However, the extralinguistic information is not restricted to data only, but may include links to exterior web pages or media files such as photographs, audio files and video clips. While the use of such additional layers of data depends on the intended use of the linguistic map, it shows that there is a broad field of applications for the display of extralinguistic information in digital mapping.

4. Prerequisites for a full digitalisation of *The Linguistic Atlas of Scotland*

4.1. Data normalisation

The representation of lexical items on a map can be understood as a model of a particular linguistic reality. In their introduction to *The Linguistic Atlas of Scotland*, Mather and Speitel (1975: 2) are very clear about the goals of their endeavour, that is "to uphold and develop a continuing and *coherent* academic discipline in linguistic geography as much as to *systematise* and publish the results of its lexical or phonological researches" [emphasis added]. In other words, their aim is to create a coherent system. In his criticism of "the Neo-positivist Doctrine of Science" (Badiou 2007: 18), Badiou claims that "the construction of a formal system [...] aims at tracing out the strict deductive structure, the mechanizable aspect, of an existent scientific domain [...]. To verify that a formal system expresses that structure well, one must bring its statements into a correspondence with the domain of scientific objects under consideration." (ibid. 19) The relation between the model and reality is crucial here. Badiou illustrates the problem by quoting an example by Rudolf Carnap, a German mathematician and proponent of logical empiricism:

> [I]f the experiment [l'expérience] can be bound to mathematical algorithms, if it is calculable, this is so insofar as phenomena can be measured. Measurement, through which facts become numbers, is here an essential semantic operation. But every result of measurement is expressed in a rational number (more precisely, a number that has only a finite number of decimals), because the 'concrete' operations of measure are necessarily finite. Semantics imposes itself on physics only as a field of numbers grounded in the field of rationals. [...] The adoption of this field as a base for physics, consequently, stems from an exigency of syntactic simplicity. (Badiou 2007: 20-21)

In the case of a linguistic survey, the analogy leads to the conclusion that already the restrictions of the questionnaire impose a limitation on the linguistic model. While the problem can be easily ignored when editing a printed version of a map, fitting informants and their responses into a database-model often requires a much more rigid approach. For example, the informant 21 from Orkney (Mather & Speitel 1975: 380) with the initals T.M.W. is male, however, the questionnaire was "[c]ompleted by several local people, all over middle age". By requiring responses to be assigned to a single, clearly identifiable individual, the questionnaire – and even more so the database-model – ignores the fact that language is always a communication process between two or more individuals. In reality, a joint effort to answer a lexical questionnaire such as the LAS-form might produce even more natural responses than those provided by an individual in a setting, in which the informants answer questions isolated from their natural language environment. As a result, what the question raises is the evaluation of 'correct' answers on the one hand, and 'incorrect', 'incomplete' or even 'too detailed' responses on the other. Thereby, it must be clarified

how the latter cases can be appropriately represented on a map, whose underlying structure systematically excludes such aberrant entries.

The data of *The Linguistic Atlas of Scotland* contains several types of responses, which do not fit readily into a database-system. As far as the informants are concerned, there are incidents in which two or more people answered the questionnaire together, either anonymously, e.g. in the case of informant "Orkney 21" (Mather & Speitel 1975: 380), or with detailed information on two or more participants, e.g. in the case of "Sutherland 3" (ibid. 381). Moreover, the date of some informants include additions to localities, e.g. "Sutherland 5", whose father was born in "Stoer, by Lairg" (ibid. 381), or the indication of half years, e.g. in the case of "Aberdeen 71b", whose length of residency is indicated with "13½" years (ibid. 385). In terms of the informants' responses, different spellings of seemingly identical lexical items are rather frequent, e.g. in the case of "*coot*, *cuit*, *cut*, *cute*, *keet*, *keut*, *kit* and *queet*" (Hessle 2019: 8). In other cases, the informants simply indicated the Standard English word given (cf. Kirk 2019: 12) or left the answer-sheet blank (cf. ibid.). A common way how to handle such aberrant incidents would be the establishment of guidelines according to which the data can be normalised, and the addition of a comment about the modification. While such comments can be easily integrated into digital maps, i.e. in form of a pop-up, the question remains whether there are better forms of representation.

4.2. Machine readability

The technology of optical character recognition (OCR) provides a good insight into the limits of automatically digitising data-lists for database-use. The most common errors are confusions of similar-looking characters, e.g. the small letters <i> and <l>, the number <1> and the capital letter <I>, or the cluster <rn> and the character <m>. Moreover, blank spaces are often not interpreted correctly. The digitisation of the data-lists of *The Linguistic Atlas of Scotland* shows that some letters are occasionally left out, for instance, the <f> indicating the 'female' sex of an informant, e.g. in the case of informant "Berwick 4" (cf. Figures 4 & 5). In other cases, several lines of the list are collapsed into a single field, in which some values are rearranged, e.g. informants "East Lothian 10-21" (cf. ibid.). In most cases, the reason for such a misinterpretation of the printed data-list is either an unexpected line-break as in the case of "Berwick 4", or a comment stretching over several columns, as with "East Lothian 10". Interestingly, the disarrangement does not only concern the respective rows, but all subsequent rows until a visual reset-indicator is identified. As a result, the

data-set provided for automatic machine reading must follow an extremely rigid structure, since already minor irregularities hold the potential to disrupt the interpretation.

32	Stow	36/4644	J.L.F.	m	64	60	Stow	same	same
EAST LOTHIAN									
1	Gullane	36/4882	W.B.	m	40	2	East Lothian	same	same
2	North Berwick	36/5585	I.W.	f	59	56	North Berwick	Auchinleck, Ayrshire	North Berwick
3	Dirleton	36/5183	W.J.S.	m	64	59	Ratho, Midlothian	West Kilbride, Ayrshire	West Kilbride, Ayrshire
4a	Cockenzie	36/3975	J.H.	m	78	74	North Shields, Northumberland	Cockenzie	Cockenzie
4b	Cockenzie	36/3975	E.M.M.	f	50	44	Kirkcaldy, Fife	Dumfermline, Fife	Armadale, Sutherland
5	Prestonpans	36/3874	J.S.	m	75	50	East Linton	North Berwick	North Berwick
6a	Tranent	36/4072	M.S.H.	f	71	68	Tranent	East Lothian	East Lothian
6b	Tranent	36/4072	P.M.	m	76	74	Tranent	same	same
7	Macmerry	36/4372	P.O.	m	68	65	Macmerry	Penston	Penston
8	Athelstaneford	36/5377	G.T.B.	m	67	64	Needless, Athelstaneford	West Fortune	Leith, Midlothian
9	Haddington	36/5173	F.F.R.	m	43	40	Haddington	Dunbar	Colchester, Essex
10	Bolton (Sch.)	36/5070	J.F.F.	f	Information given by man over seventy – native who has lived all his life in district.				
11	Tynninghame (Sch.)	36/6079	E.C.G.	f	Information given by several local residents.				
12	West Barns	36/6578	J.D.	m	47	14	Kilbirnie, Ayrshire	Kilbirnie, Ayrshire	Dalry, Ayrshire
13	Dunbar	36/6778	C.G.L.D.	f	42	39	Dunbar	Ayton, Berwickshire	Dunbar
14	Spott	36/6775	W.S.	m	54	54	Spott	Spott	——
15	Whittingehame	36/6073	C.B.A.	m	66	4	Bolton, Haddington	Muthill, Perthshire	Yetholm, Kelso, Roxburghshire
16	Innerwick	36/7273	T.D.H.	m	82	79	Innerwick	Spott parish	Norham, Northumberland
17	Oldhamstocks	36/7370	G.D.	m	79	41	Coldstream, Berwickshire		
18	Ormiston	36/4169	A.S.	m	67	11	Haddington	Whitekirk	Haddington
19	Pencaitland	36/4468	W.Y.	m	46	43	Pencaitland	same	same
20	East Saltoun	36/4767	P.F.P.	m	59	12½	Wishaw, Lanarkshire	Lanarkshire	Wishaw, Lanarkshire
21	Humbie	36/4662	A.C.	f	63	12	Ormiston, Midlothian	Norham, Northumberland	Crichton, Midlothian
BERWICK									
1	Cockburnspath	36/7771	E.H.	m	42	23	Cockburnspath	Gavinton	Chirnside
2	Cranshaws (Sch.)	36/6961	E.A.	f	47	16	Duns	Galashiels, Selkirkshire	Duns
3	Abbey St Bathans	36/7662	A.I.O.	f	62	4	Strichen, Aberdeenshire	Gartly, Aberdeenshire	Strichen, Aberdeenshire
4	Greenwood	36/8304	M.I.D.W.	f	52	52	Cockburnspath	Berwickshire	Berwickshire
5	Greenhead	36/7561	N.S.	m	67	64	Greenhead	——	——

Figure 4: Data-list from LAS (Mather & Speitel 1975: 400)

Loth ian
32 Stow 36/4644 J.L.F. m 64 60 Stow same same
EAST LOTHIAN

1	Gullane	36/4882	W.B.	m	40			same	same
2	North Berwick	36/5585	1.W.	f	59			Auchinleck , Ayrshire	North Berwick
3	Dirleton	36/5183	W.J.S.	m	64			West K ilbride ,	West Kilbride,
4a	Cockenzie	36/3975	J.H.	m	78			Ayrshire Cockenzie	Ayrshire Cockenzie
4b	Cockenzie	36/3975	E.M.M.	f	50			Dumfermline, Fife	Armadale, Sutherland
5	Prestonpans	36/3874	J.S.	m	75			North Berwick	North Berwick
6a	Tranent	36/4072	M.S.H.	f	71			East Lothian	East Lothian
6b	Tranent	36/4072	P.M.	m	76			same	same
7	Macmerry	36/4372	P.O.	m	68			Penston	Penston
8	Athelstaneford	36/5377	G.T.B.	m	67			West Fortune	Leith, Midlothian
9	Haddington	36/5173	F.F.R.	m	43			Dunbar	Colchester, Essex

10 Bolton (Sch.) 36/5070 J.F.F. f Information given by man over seventy - native who has lived all his life in dist r ict .
11 Tynninghame(Sch.l 36/6079 E.C.G. Information given by several local residents .
12 West Barns 36/6578 J.D. m 47 14 Kilbirnie, Ayrshire Kilbirnie, Ayrshire Dairy, Ayrshire
13 Dunbar 36/6778 C.G.L.D. f 42 39 Dunbar Ayton , Berwickshire Dunbar
14 Spott 36/6775 W.S. m 54 54 Spott Spon
15 Whittingehame 36/6073 C.B. A. m 66 4 Bolton, Haddington Muthill, Perthshire Yetholm, Kelso, Roxburghshire
16 Innerwick 36/7273 T.D.H. m 82 79 Innerwick Spott parish Norham, Northumbe1 land
17 Oldhamstocks 36/7370 G.D. m 79 41 Coldstream, Berwickshire
18 Ormiston 36/4169 A.S. m 67 11 Haddington Whitekirk Haddington
19 Pencaitland 36/4468 W.Y. m 46 43 Pencaitland same same
20 East Saltoun 36/4767 P.F.P . m 59 12½ Wishaw, Lanarkshire Lanarkshire Wishaw, Lanarkshire
21 Humbie 36/4662 A.C. f 63 12 Ormiston, Midlothian Norham, North- Crichton, Midloth ian um berland

BERWICK

1	Cockburnspath	36/7771	E.H.	m	42	23			Chirnside
2	Cranshaws (Sch.)	36/6961	E.A.	f	47	16			Duns

3 Abbey St Bathans 36/7662 A.I.0 . f 62 4 Strichen, Aberdeensh ire Gartl y, Aberdeen-sh i re Strichen, Aberdeen-shire

4	Greenwood	36/8304	M.I.D.W.		52	52			Berwickshire
5	Greenhead	36/7561	N.S.	m	67	64			

Figure 5: OCR-Output from LAS-data

5. Conclusion

This study shows that it is possible and necessary to establish guidelines according to which the digitalisation and reinterpretation of historical data collections may lead to new insights. Thereby, the possibility to visually contextualise the linguistic data by combining it with extralinguistic information enables linguists to derive new results. Such results necessarily reflect back on the categorisation of the data-set. As the categorisation and systematisation is at the core of the interpretation of the data, this paper argues that the semantic process must remain open throughout the whole research-process. Furthermore, it becomes apparent that processes whose nature is often perceived as purely technical, must necessarily be questioned in the course of the digitisation process. In particular, questions concerning the visualisation process such as the choice of a colour-palette or, with regards to data normalisation, the evaluation of 'correct' or 'false' data entries must be treated with great caution. This is even more so required, given the narrow limits of digitisation technologies.

6. Acknowledgement

We are grateful to the permission given to John Kirk by Croom Helm to digitise the atlas maps and data.

7. References

Allred, S. Colby; Schreiner, William J.; Smithies, Oliver. 2014. Colour blindness. Still too many red-green figures. *Nature* 510, 340.

Amadae, Sonja M. 2015. *Prisoners of Reason. Game Theory and Neoliberal Political Economy.* Cambridge: Cambridge University Press.

Bejinariu, Silviu-Ioan; Olariu, Florin-Teodor. 2017. Romanian Linguistic Atlases in Digital Format – A New Approach. *Philologica Jassyensia* Vol. XIII(1/25): 13-23.

Badiou, Alain. 2007 [1969]. *The Concept of Model: An Introduction to the Materialist Epistemology of Mathematics.* Melbourne: re.press.

Badiou, Alain; Haéri, Gilles. 2016. *In Praise of Mathematics.* Cambridge: Polity Press.

Barthes, Roland. 2005 [1977-1978]. *The Neutral. Lecture Course at the at the College de France (1977-1978).* New York: Columbia University Press.

Flegg, Henry G. 1974. *From Geometry to Topology.* London: The English University Press.

GeoJSON. 2019. The GeoJSON Format https://tools.ietf.org/html/rfc7946 (19 Jan. 2019).

Hessle, Christian. 2019. *Towards A Digital Version of The Linguistic Atlas of Scotland*. Unpublished BA Dissertation, University of Vienna.

Johnston, Paul. 1997. Regional Variation. In Jones, Charles (ed.). 1997. *The Edinburgh History of the Scots Language.* Edinburgh: Edinburgh University Press, 443-513.

Kirk, John M. 1994a. East Central Scots: A Computerised Mapping Package. In Fenton, A.S. and MacDonald, D.A. (eds.). *Studies in Gaelic and Scots.* Edinburgh: Canongate Academic, 48–68.
Kirk, John M. 1994b. Maps: Dialect and Language. In Asher, R.E. Asher and Simpson, J.M.Y. (eds.). *The Encyclopedia of Language and Linguistics. Volume 5*. Oxford: Pergamon Press, 2363–2377.
Kirk, John M. 2019. On the Digitisation of Lexical Survey Material. Unpublished presentation at the *Scots Words and Phrases in the Contemporary World: Back to the Future* Symposium, University of Edinburgh, 8-9 April.
Kretzschmar, William A. 1992. Isoglosses and Predictive Modelling. *American Speech* 67(3): 227-249.
Leaflet. 2019. Leaflet – a JavaScript library for interactive maps. https://leafletjs.com/index.html (19 Jan. 2019).
Leck, Ian. 1994. Is a minor inconvenience. *British Medical Journal* 309(6947), 129.
Little, William; Onions, Charles T. (eds.). 1956. *The shorter Oxford English Dictionary on historical principles.* (3rd edition). Oxford: Clarendon Press.
Macaulay, Ronald K.S. 1977. Review of Mather, James Y.; Speitel, Hans-Henning (eds.). 1975. *The Linguistic Atlas of Scotland: Scots Section. Volume 1.* London: Croom Helm Ltd.. *Language* 53: 224-228.
Macaulay, Ronald K.S. 1979. Review of Mather, James Y.; Speitel, Hans-Henning (eds.). 1977. *The Linguistic Atlas of Scotland: Scots Section. Volume 2.* London: Croom Helm Ltd.. *Language* 55: 224-228.
Macaulay, Ronald K. S. Linguistic Maps: Visual Aid or Abstract Art? In Kirk, John M.; Sanderson, Stewart; Widdowson, J. D. A. 1985. *Studies in Linguistic Geography. The Dialects of English in Britain and Ireland*. London: Croom Helm, 172-186.
Mapbox. 2018. About Mapbox https://www.mapbox.com/about/ (19 Jan. 2019).
Mather, James Y.; Speitel, Hans-Henning (eds.). 1975. *The Linguistic Atlas of Scotland. Scots Section. Volume 1.* London: Croom Helm Ltd.
Mather, James Y.; Speitel, Hans-Henning (eds.). 1977. *The Linguistic Atlas of Scotland. Scots Section. Volume 2.* London: Croom Helm Ltd.
McClure, J. Derrick. 1976. The Linguistic Atlas of Scotland. *American Speech* 51(3-4): 223-234.
Murison, D.D. 1978. Review of Mather, James Y.; Speitel, Hans-Henning (eds.). 1977. *The Linguistic Atlas of Scotland: Scots Section. Volume 2.* London: Croom Helm Ltd.. *The Scottish Review* 9: 45-48.
OpenStreetMap. 2018. History of OpenStreetMap. https://wiki.openstreetmap.org/wiki/History_of_OpenStreetMap (7 Dec. 2018).
Ordnance Survey. 2019. https://www.ordnancesurvey.co.uk/business-and-government/counties/index.html (23 May 2019).
Price, Matthew. 2013. Crowsteps in Fife: The Flemish Connection, Part 1. *Scotland and the Flemish People Project.* St. Andrews: St. Andrews University. https://flemish.wp.st-andrews.ac.uk/2013/12/06/crowsteps-in-fife-the-flemish-connection-part-1/ (10 Jan 2019).
Scottish Language Dictionaries (ed.). 2017. *Concise Scots Dictionary* (2nd edition). Edinburgh: Edinburgh University Press.
Scottish Language Dictionaries (ed.). 2018a. *Dictionary of the Older Scottish Tongue Electronic Version.* http://www.dsl.ac.uk (20 Dec. 2018).
Scottish Language Dictionaries (ed.). 2018b. *Scottish National Dictionary Electronic Version.* http://www.dsl.ac.uk (20 Dec. 2018).
Xie, Yuchun; Aristar-Dry, Helen; Aristar, Anthony; Lockwood, Hunter; Thompson, Josh; Parker, Dan; Cool, Ben. 2009. Language and Location: Map Annotation Project - A GIS-Based Infrastructure for Linguistics Information Management. *Proceedings of the 2009*

International Multiconference on Computer Science and Information Technology. 305-311. [downloadable from https://fedcsis.org/archive/proceedings_2009]