Journal of Data Mining and Digital Humanities

Digitising Collections of Historical Linguistic Data:

The Example of The Linguistic Atlas of Scotland

Christian Hessle, John Kirk*

University of Vienna, Austria

*Corresponding author: John Kirk john.kirk@univie.ac.at

Abstract
This article addresses the issue of variation in the lexicon – specifically the hyponymy (or synonymy) among onomasiological responses for the same concept or referent – and how the range of responses from a national elicitation in Scotland seeking 'local' words should be judged. How do responses being offered as 'local' square with their geographical distribution on the one hand, and their status as 'Scots' or 'English', or as 'dialect' or 'standard' on the other? How are 'dialect' or 'standard' responses offered as 'local' responses from the same individual to be considered? Is the issue that of a straightforward dialect-standard binary opposition, or is there a third value between the two? Does that third value encompass a middle ground between dialect and standard, or include both? How is the absence of responses to be regarded? For elucidation of such linguistic issues, the article invokes the mathematical principle of the excluded middle. This study shows that it is possible and necessary to establish a theoretical framework for the digitalisation of a historical data collection.
The data for these reflections come from the lexical material in *The Linguistic Atlas of Scotland* ([Mather and Speitel, 1975]; [1977]), which is currently being digitised at the University of Vienna. This study presents three pilot studies from the North Mid Scots area: the atlas concepts of 'ankle', 'splinter' and 'the youngest of a brood or litter'. The original data are re-analysed in terms of lexical types or 'lexemes', and the results are digitally generated by new dot maps, with separate gestations for respondent age and gender. In the process, topological issues (such as those pertaining to the data) and topographical issues (such as those pertaining to geography and the physical terrain) are addressed.

keywords
Scots, lexis, dialect, standard, local, maps, cartography, binary oppositions, the excluded middle

INTRODUCTION

The age of digitalisation opens up new perspectives on linguistic geography. 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 and Speitel, 1975]; [1977]). After an introduction to the atlas, covering its contents and design as well as its shortcomings, a prototype project 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 presuppositions 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 historical data collections can be digitally processed and thus touches on the limitations of data normalisation and machine readability.

Our data qualifies as historical in several respects. That the data were collected in the 1950s and 1960s and are already in themselves a fixed historical object is not the main point. Rather, for the concepts a great many of which refer to human beings, their bodies and their characteristic features, as well as to the natural world, to insects, beasts and farm animals, and to the land and ways of farming the land, those data have clearly been in oral currency for centuries. Moreover, their geographical distribution inevitably aligns with historical patterns of language distribution. For us, the considerable variation in the lexical data for each concept, on each map, can only properly be explained historically. It is the etymological record of the atlas data which emphasises the diachronic character of the lexical items for long-standing concepts, and which substantiates our historical perspective on the data. On our maps, geography and history become combined. But there is a further point: because of social change, through the mechanisation of farming, and also through urban residence, many of those words have undergone erosion, obsolescence or downright loss (cf. e.g. [Upton and Widdowson, 1999]). The interactivity of an online atlas has the potential to make these data accessible in ways like never before and consequently to revive knowledge of such words and the concepts to which they have traditionally related, not least as an educational resource, for both linguistic as well as wider cultural studies.

I. 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 and 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) in the fields of the human body, clothes, agriculture, animals and children's games. All in all, the first volume of the atlas includes responses by 1,774 informants ([Mather and Speitel 1977: 9]). The results are presented on 122 linguistic maps and lists for 90 lexical items ([Mather and Speitel 1975: Contents]), providing a detailed illustration of lexical items on a concept-by-concept, map by map, basis throughout Scots-speaking areas. 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.
Figure 1 shows the published map for 'splinter' ([Mather and Speitel, 1975: Map 4]).

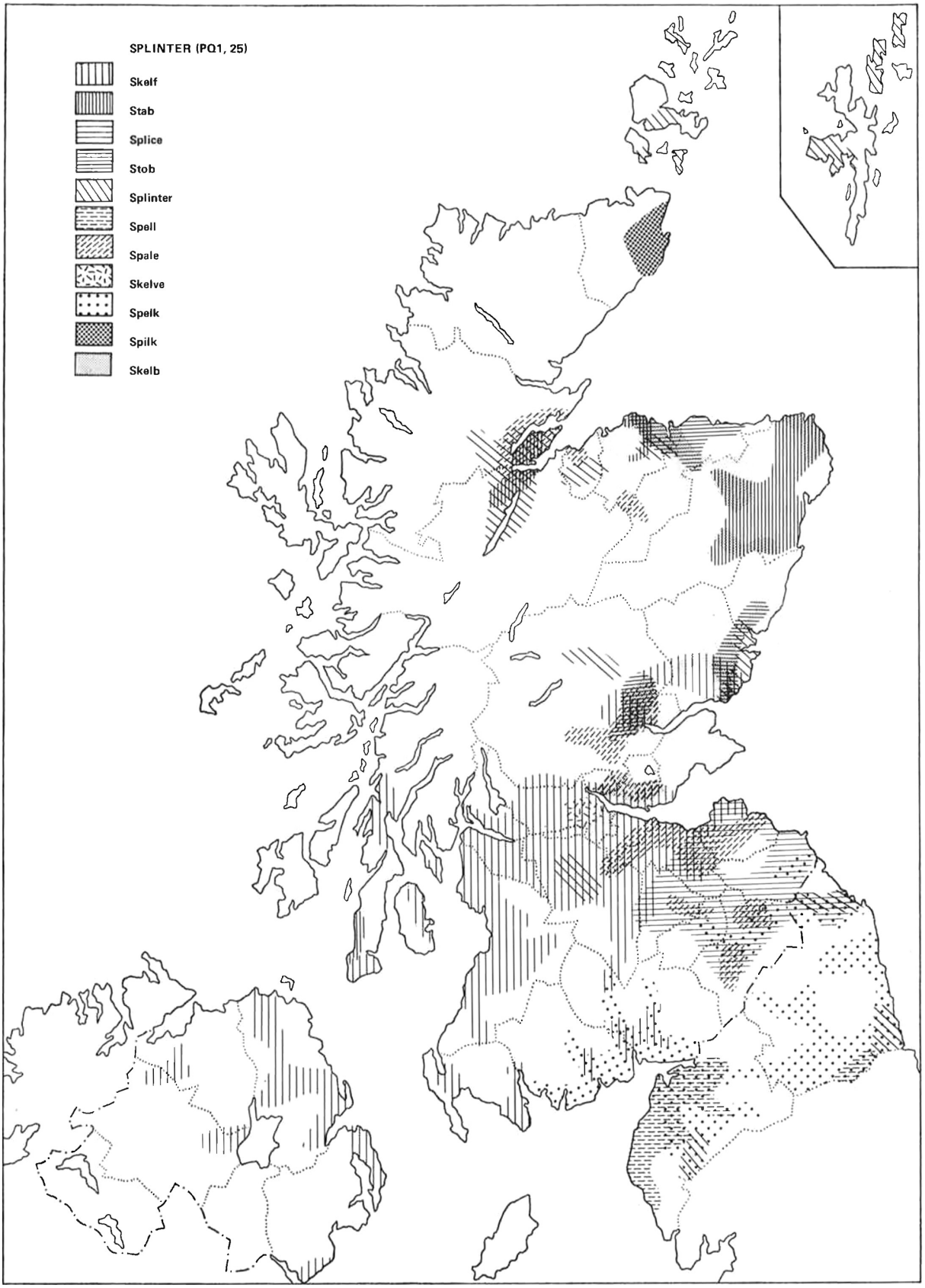

Figure 1. *The Linguistic Atlas of Scotland*, map for 'splinter' from [Mather and Speitel, 1975: Map 4]

Many will concur with [Derrick 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 from 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", indicated as 'Nil' response in the atlas's data-lists. 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) as if separate responses or 'words'. For example, in the map for 'Splinter' (cf. [Mather and Speitel, 1975: 33], reproduced here as Figure 1), the items 'Skelve' and 'Skelf' are indicated as separate items, while, as analysis shows, both items belong to the same lexical group, rendered as *skelf* (cf. [Hessle, 2019: 13]). On the other hand, hapax items that are "attested only once" ([McClure, 1975: 230]) are generally not represented on the maps, although many of these hapaxes are simply further orthographic variants of words which are indeed mapped. An earlier 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 or variants. [Ron Macaulay, 1985: 175] acknowledges 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 inconsistency 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, we suppose others could 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 and 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", which the digitised atlas will be showing in abundance, as does the prototype ([Hessle 2019]).

II. TOWARDS A DIGITAL VERSION OF *THE LINGUISTIC ATLAS OF SCOTLAND*

The unpublished study entitled *Towards a Digital Version of The Linguistic Atlas of Scotland* ([Hessle, 2019]), which includes three of the questionnaire items, namely 'ankle', 'splinter', and 'youngest of a brood', is restricted to 182 informants from pre-1975 Scottish counties of Clackmannan, Fife, Kinross and Perth, comprising 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 interactive

maps for the three lexical items (cf. [Hessle, 2019]). 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. By providing an outlook on how a future digital version of *The Linguistic Atlas of Scotland* (LAS) might be realised, Hessle's study also identifies a number of challenges which might arise during the process of digitisation, to be dealt with here under the headings of *lexical analysis*, *issues of categorisation*, and *the mapping process*.

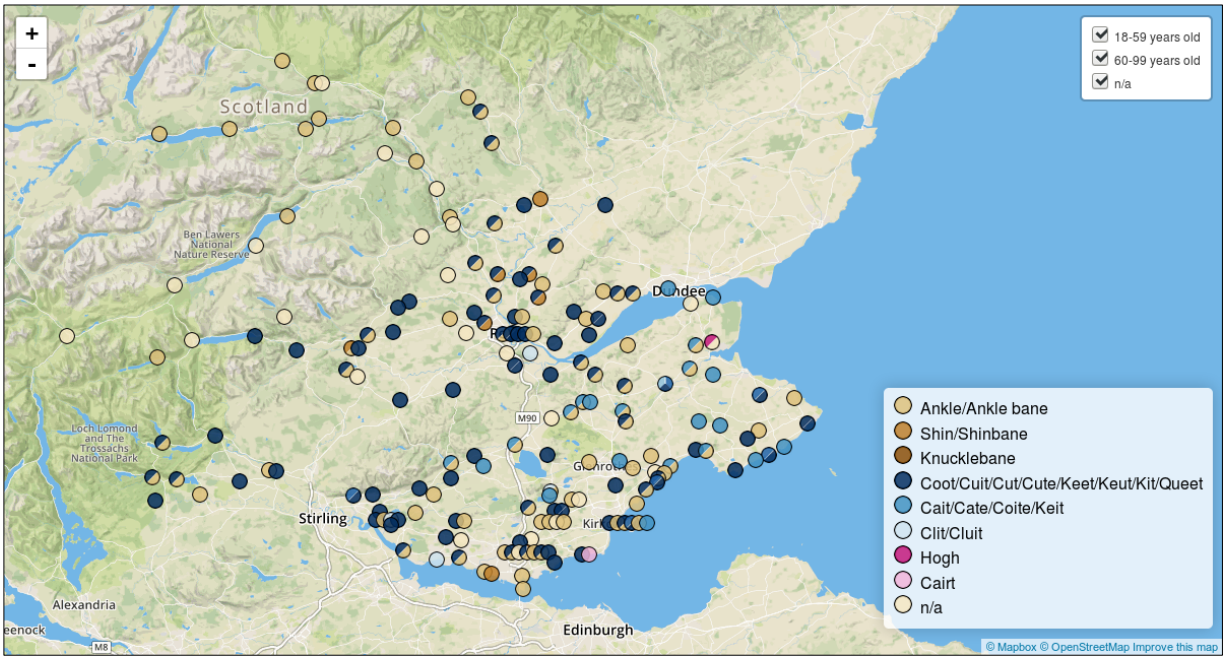

Figure 2. Scots synonyms for SSE 'ankle' (sorted according to age groups) in Clackmannan, Fife, Kinross & Perth (Map 1 from [Hessle, 2019], accessible online at http://16levels.org/las/ankle_age.html)

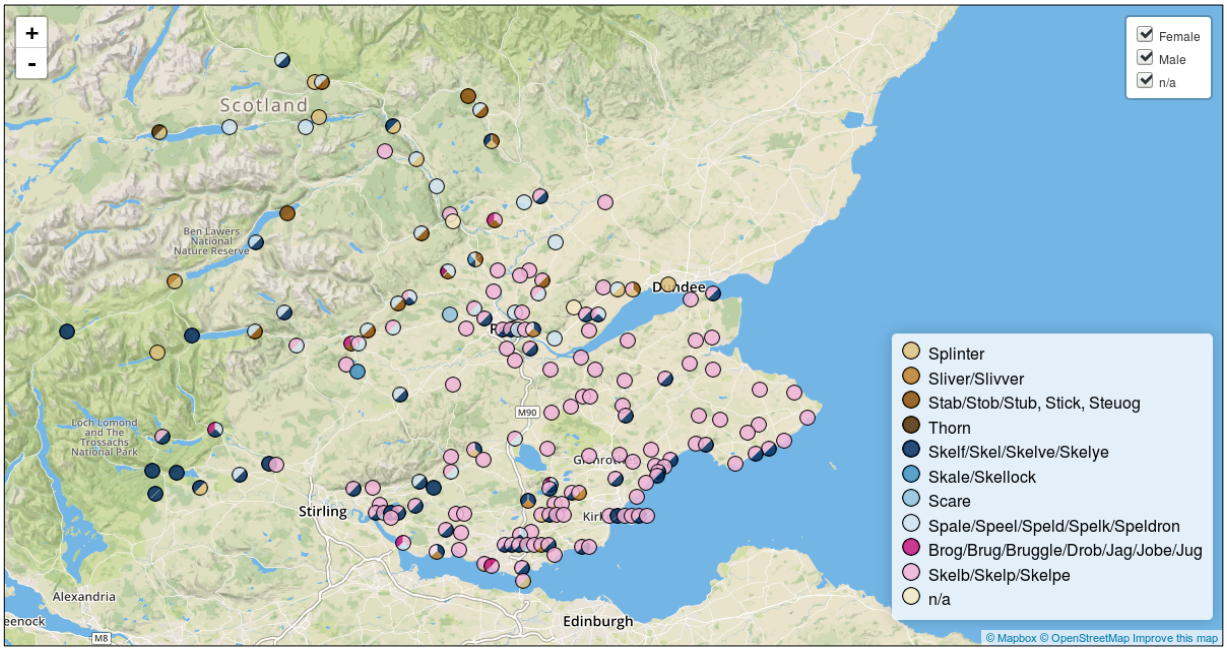

Figure 3. Scots synonyms for SSE 'splinter' (sorted according to gender) in Clackmannan, Fife, Kinross & Perth (Map 2 from [Hessle, 2019], accessible online at http://16levels.org/las/splinter_gender.html)

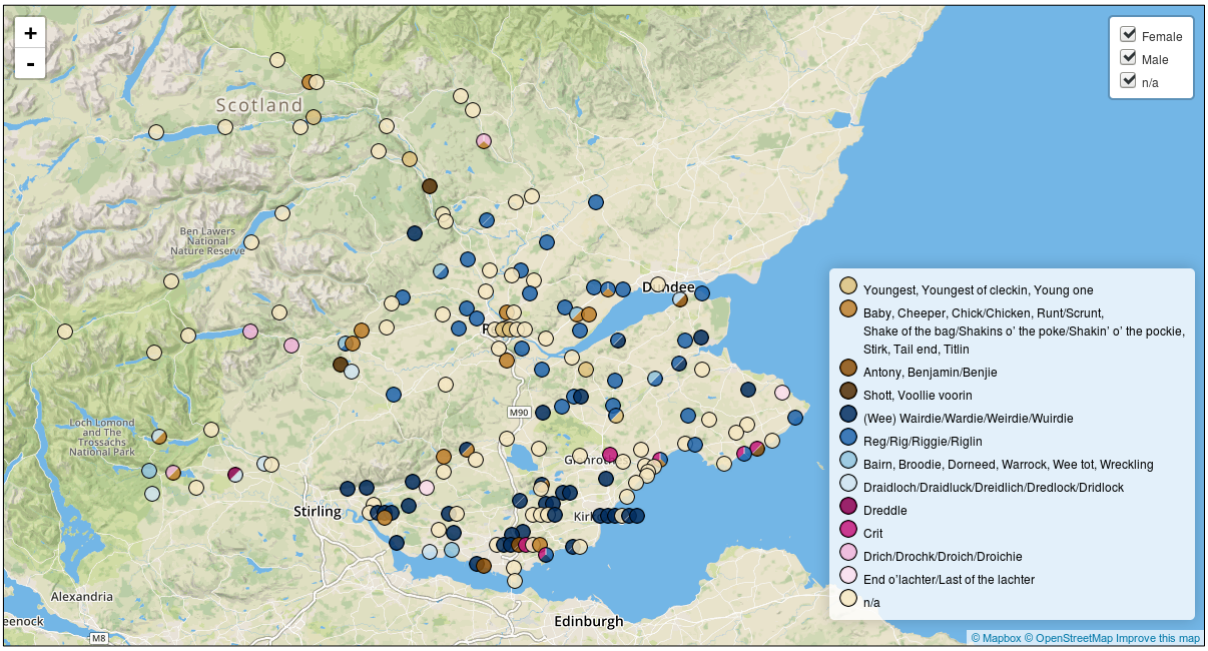

Figure 4. Scots synonyms for SSE 'youngest of a brood' (sorted according to gender) in Clackmannan, Fife, Kinross & Perth (Map 3 from [Hessle, 2019], accessible online at http://16levels.org/las/youngest_gender.html)

2.1. Lexical analysis

For each concept, LAS provides an abundance of responses. A great many responses are simply orthographic variants of a lexical item, all listed in the atlas. This array of orthographic types arises from the request in the original questionnaire for respondents to write down, in any way they thought appropriate, their usual or local pronunciation of the word. In a great many cases, there was no conventional way for doing so, so that informants simply had to make up the spellings as they saw fit. We, however, are concerned neither with orthographic variants as such, nor the pronunciations which they are supposed to represent; rather, we are only concerned with lexical types – different words or, more strictly, lexemes – as occasionally the response is a lexical phrase rather than a single word. For each range of orthographic variants, there is usually a recognised spelling of the head item or *hypernym*, but where necessary a hypernymic spelling has had to be chosen on the basis of the head word spelling in the *Concise Scots Dictionary*, or from the second author's knowledge of Scots. In identifying separate lexemes, we will make use of all the etymological information that we can; in some cases, cultural practices or patterns of belief, such as paintings and publications in the fields of architecture and history, will also be taken into account. Such references will help identify cases of semantic transfer. Last but not least, in order to identify different orthographic forms of the same lexical item, Paul Johnston's [1997] word-sets model is used as a further basis for categorising the informants' responses. This process of establishing lexemes we will call 'lexemisation'.

For the etymological part of the analysis of Scots, we will rely on the *Concise Scots Dictionary* [2017] (hearfter, CSD) as well as 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). For English, including both historical and geographical varieties, we will consult the online *Oxford English Dictionary* and the related *Shorter Oxford English Dictionary*

As an example of the cultural-historical sources we used, [Hessle, 2019: 8] shows that among the responses given for the concept of 'ankle', the orthographic variants *cait*, *cate*, *coite* and *keit*, which refer to the game of curling, are attested. While the connection between *ankle* and *curling* remains unclear in the dictionary 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.). And hence, from this cultural practice, still current in Saskatchewan today (cf. [Hill, 2016]), we deduce a clear case of semantic transfer.

It is well-known that, at the invitation of King David I, Flemish settlers arrived in Scotland in the 12th century. Many of them settled on the southern shore of the Firth of Forth, in places such as Culross, where signs of Dutch influence, such as the "typical crow-stepped gable[s]" ([Price, 2013]) can still be seen today. It is not surprising that it is in areas of Flemish settlement that words from Middle Dutch tend to be found. It is also clear that Scottish Gaelic was spoken in Fife until well into the eighteenth century. The predominance of Gaelic in Fife is shown by the great many place names of Gaelic origin, a striking example of which is the former name of St. Andrews: *Kilrymont* (literally "the church of the king on the hill').

Throughout the North Mid Scots area, some individual responses include *skelp* and others *skelve*, which we interpret simply as orthographic variants of *skelb* and *skelf* respectively indicative of a contrast of voicing in the final consonant. But in Fife, especially along the southern coast, some respondents give both *skelf* and *skelb*. While *Skelf* is probably a loan from Middle Dutch *schelf*, ('a flake', 'a splinter of wood', according to the CSD), *skelb* is a loan from Scottish Gaelic *sgealb* ('a splinter'', 'a flake', according to the CSD)- quite separate lexemes on etymological grounds, and not at all a question of phonological variants resulting from final consonant (de)voicing. Moreover, the fact that some respondents gave both *skelf* and *skelb* suggests that they were able to distinguish between the two lexical groups not as alternatives of the same lexeme but as different lexemes. Thereby, it shows that informants are frequently able to clearly distinguish between the two lexical groups *skelf* and *skelb*. Accordingly, we interpret *skelb* and *skelf* as separate lexemes, each a Scots synonym for Standard English *splinter* ([Hessle, 2019: 13]), *CSD* notwithstanding.

2.2. Categorisation

As already explained, the responses elicited by the original questionnaires ranged from standard words to dialect words; from phrasal descriptions, to evaluations and metaphorical extensions; from words that denoted the concept directly, to those which did not, or which displayed semantic transfer (such as words for other parts of the leg being given in response to 'ankle', see below). In this section, we attempt a fresh theorisation of this range of responses. We shall see that the multiple responses given for each concept require modelling beyond a simple binary choice between X and not-X, such as standard vs. non-standard (dialect).

The twentieth century saw the emergence of multi-valued logical systems, which – in contrast to classical logic – do not adhere to the *principle of the excluded middle*. In questioning the foundational principle of classical mathematics "that the negation of negation is equivalent to simple affirmation," ([Badiou, 2016: 60]) *n*-valued logical systems are "not restricting the number of truth values to only two, thus allowing for a larger class of '*degrees of truth*'" ([Georgescu, 2006: 123]). In 1920, Polish mathematician Jan Lukasiewicz introduced a three-valued logic entitled $Ł_3$, "in which propositions may be classed as True, False, or Uncertain, the third value being the excluded middle of the traditional two-valued system." ([Leonard, 2010: 126]). Lukasiewicz's approach was connected to the intuitionistic logic of Luitzen E. J. Brouwer, who was working at that time on how set theory could be applied to topology – what

Graham later calls "the study of non-metric spatial relationships [and their] continuity" ([Flegg, 1974: 19]). During the 1930s, proponents of logical empiricism such as John von Neumann and Oskar Morgenstern continued to base their theoretical frameworks on a binary logical system which, according to [S. M. Amadae, 2015: 26] "captured the mentality of competitive duelling exercised in parlour games and military combat." In their game-theoretic concept, for instance, social interaction is simplified to what Amadae calls a "two-person 'game'" ([2015: 26]) in which each participant holds only two options, before being extended to "a multi-person, indefinitely repeating" (ibid.) scenario. By these accounts, communication and language are reduced to a mere "signalling game" (ibid: 39) deprived of meaning. More recently, such a universalist approach to logic has been criticised by the philosopher-mathematician, Alan Badiou, who argues that topology follows an intuitionistic logic ([Badiou, 2016: 62]). For linguistic topology, Badiou contends, it is essential to "define the rules of correspondence. Everything concerning these rules depends on the *semantics* of the system, on its *interpretation.* [...] to speak of the meaning of the system is to speak of its various interpretations" ([Badiou, 2007: 19]).

For us, intuitionistic logic – unlike a binary system adhering to the principle of the excluded middle – serves as an appropriate framework for linguistic topology (i.e. linguistic variation). As the elements in the system, our topology comprises the lexemes which have arisen from the lexemisation of the orthographic responses provided by the questionnaire. It is those lexemes which we are proceeding to add cartographically to each map, using symbolised encodings representing each lexeme. By contrast, what the map (for more details, see below) represents is the topography of Scotland: its coast, its rivers, its lakes, its mountain and hill ranges, its cities, and so on. Thus, between the physical, topographic map and the cartography superimposed on it, there lies for us the intrinsic value of the atlas: the challenge of interpreting the status and relationship of lexemes to each other within a system (topology), but at the same time also of interpreting their relationship to – and distribution across – the land-geography of Scotland (topography). The former is concerned with the status of responses: whether they are standard or dialect, or whether they are what we are calling 'denotans' or 'non-denotans', or whether they are descriptions, evaluations and metaphorical extensions, etc. and what the linguistic relationship between them might be; the latter is concerned with the location of those responses in relation to their provenance, as revealed by the map. In these ways, a clear distinction between the physical map and the cartography superimposed on it becomes apparent. Expressed as a research question, our goal may be formulated as follows: what is the relationship between the topological and topographical distribution of the lexemes gathered by the LAS? Or less formally, what is the relationship between the lexical and geographical distribution of the lexemes gathered by the LAS? We suspect that this formulation would not have been far from the aims and objectives of the original compilers.

While the analysis of the LAS-data makes apparent the contradiction between language and a binary logic regime, it also shows that a descriptive approach on the basis of etymology and semantics provides a set of linguistic categories to account for 'the excluded middle' and all its diversity. In asking for "word or words commonly used for [concepts expressed in Standard English] in [their] own locality", [Mather and Speitel, 1975: 13] were prescribing a dichotomy between Scots and Standard English at the outset. But the answers to their questionnaire abundantly show that such a binary choice does not hold. On the one hand, 'nil'-answers indicate that there must be at least a third set of responses corresponding to uncertain values in *n*-valued logical systems. As already mentioned, McClure's [1975: 227] suggestion of "three possible interpretations of an informant's failure to respond to an item in the questionnaire" is a clear indication of such an uncertain value. On the other, there are many cases in which the

answers for Scots-items has the same status as the given Standard English item; and there are many other cases in which both Scots and the Standard English lexeme are given – for example, in the cases of responses for the item 'ankle' in Kinross-shire and Clackmannanshire:

Respondent 1	*Kit*
Respondent 2	*ankle, cait*
Respondent 3	*ankle, cait*
Respondent 4	*Cait*
Respondent 5	*Cuit*
Respondent 6	*ankle, cuit*
Respondent 7	*Cuit*

Table 1. *The Linguistic Atlas of Scotland*, responses for 'ankle' in Kinross-shire ([Mather and Speitel , 1975: 153])

Respondent 1	*clit, cuit*
Respondent 2	*Cuit*
Respondent 3	*Kit*
Respondent 4a	*Kit*
Respondent 4b	*ankle*
Respondent 4c	*ankle, clit*
Respondent 4d	*Kit*
Respondent 5	*Kit*
Respondent 6	*ankle*
Respondent 7	*Kit*

Table 2. *The Linguistic Atlas of Scotland*, responses for 'ankle' in Clackmannanshire ([Mather and Speitel, 1975: 153])

In our atlas, responses *cait* and *cuit* will be lexemised to *cuit*. Tables 1 and 2 show that in Kinross-shire three respondents have the standard as well as the Scots word, and four have only the Scots word, whereas in Clackmannanshire two respondents have only the standard word, one has the standard and a Scots word, and seven have only Scots words. These simple examples are sufficient to prove that a model based on intuitionistic logic, allowing for both/and as well as neither/nor possibilities, is what is required to explain the inherent variation in the data. After all, in the North-Mid Scots area, this category includes other Scots words such as *hogh* (which is given only by respondent Fife-4, who gives *ankle* as well), *knucklebane*[1] and *shin/shinbane*.[2] Such a range of responses does not fit the classical logic system, where an item can never be both true and false at the same time, but rather that it is the "larger class of '*degrees of truth*'" (original italics [Georgescu 2006: 123]) in a multi-valued logic which allows us to understand the linguistic topology of Scotland. This approach presents us with a language continuum as opposed to an imagined, idealised binary language environment. Our contention is that the lexicology of Scots is simply not a binary Scots vs. English set of alternative, mutually-excluding choices. Moreover, that range of responses – the many third and other responses of the unignorable middle ground – further confirms that prescriptive modelling of linguistic topologies – drawing on strategies seeking the establishment of a universal model through a radical simplification in combination with infinite repetition – is not an appropriate

[1] Although the noun *knuckle* is often associated with fingers, it is defined by the CSD §4 as 'the protuberant bone at the joint of the knee, wrist or ankle', thus obviously associated with the ankle by the one respondent (Fife-55d).

[2] *shin* is defined in the CSD as 'the front part of the leg between the leg and the ankle'; and as 'the edge of the shank bone', thus obviously transferred to the ankle itself by the seven respondents.

approach to linguistic reality. Hence, for us, a descriptive approach based on etymology and semantics seems well suited to outline a linguistic topology of Scotland.

Our contention is reinforced by lexicography: specifically the *CSD* and the *Scottish National Dictionary* (SND). If an explanation of the middle ground entails 'uncertainty' (as advocated by Lukasiewicz), such uncertainty lies behind the treatment of many items in the dictionary. Thus, as a response for 'ankle', *hog* [sic] is described as "unknown; perhaps related to HOGG[1.1]" [CSD: 302] (see Figure 5) and that "its relationship to Eng. *jog* is uncertain" (SND) (see Figure 6), thereby indicating a non-binary logic.

> **hog**[1.1] /hɔg/ *n curling* a stone which does not pass over the hog score *la18-*. **hog score** the lines at each end of the rink over which a shot must pass to score *la18-*. [unknown; perhaps related to HOGG[1.1]]

Figure 5. *Concise Scots Dictionary*, entry for *hog*[1.1] (CSD: 302)

> **JOG**, *v., n.* Also *joog, joug, jug(g), jowg*. [Sc. ʤog; Per. ʤug. Fif. ʤʌug]
> **I.** *v.* 1. To prick, pierce with a sharp instrument (Sc. 1710 T. Ruddiman GL to Douglas *Aeneis*; Gall. 1824 MacTaggart *Gallov. Encycl*. 281; Uls. 1880 Patterson *Gl*; Dmf. 1925 *Trans. Dmf. & Gall. Antiq. Soc*. 30; Fif. 1926 Wilson *Cent. Scot*. 250; Kcd. m.Sc.., Rxb., Uls. 1959) *Cf.* JAG
> Per 1898 C. SPENCE *Poems* 72:
> Lang ha'e I trod in folly's path, Sair jogged wi' thorns and nettles scaudie.
> Per[2] 1928
> I've joogit ma thoomb.
> **2**. To have sexual intercourse with (a woman)
> Sc. 1736 Slang of Venery (1916) 1. 159:
> Tho' he jog'd me sprightly.
> **II.** *n.* A prick, a jab with something sharp (Fif. 1926 Wilson Cent. Scot. 250; Kcd., Ags. 1959)
> [A variant of JAG, *n.*[1], v[1], of imit. orig. its relationship to Eng. *jog* is uncertain.]

Figure 6. *Scottish National Dictionary*, entry for *jog* (accessible online at https://dsl.ac.uk/entry/snd/jog)

Lexicographical uncertainty thus confirms lexicological uncertainty, including an item's status as dialect or standard, Scots or English, although mere inclusion in the dictionary inevitably preferences an item's Scots/dialect status.

To return to our research question (what is the relationship between the lexical and geographical distribution of the lexemes gathered by the LAS?), an interpretation of the linguistic topography of Scots depends on the quality and certainty of the linguistic categories on which the cartography comes to be based – the degree of certainty entailing the degree of proper topographical contextualisation.

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

Let us now consider the digital technologies which were used in the pilot study for mapping the data of *The Linguistic Atlas of Scotland* (cf. [Hessle 2019]). As the base-map, the Open Street Maps-project (cf. [2018]) was chosen. The open source project founded by Steve Coast was initially focussed on mapping the United Kingdom, which ensures the necessary topographical detail. Since the grid provided for the informants' localities in *The Linguistic Atlas of Scotland* does not 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 linguistic information cartographically, we decided to use the publicly available 'tiles' for topographical information provided by Mapbox (cf. [2019]), which could be layered on the maps (cf. [Hessle 2019: 7]), as shown in Figures 2-4 above.

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 be run locally in a web-browser. Furthermore, the technologies used are available under an open source license and include extensive documentation. Even more importantly, 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 need to constantly 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. Figures 2, 3 and 4, 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 were taken in order to optimise the readability of the study's maps. As Figures 2, 3 and 4 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 (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 *et al.* 2004] 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[3] "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 all-over organization [...] that in a way skips the

[3] In Hessle's map, 'Nil' responses are labelled "n/a" (cf. Figs. 2, 3 and 4)

paradigm" (ibid.) of the semantics of the visual representation of the linguistic data. Barthes suggests to think of these nuances as a *moiré-pattern* "whose aspect, perhaps whose meaning, is subtly modified according to the angle of the subject's gaze." (ibid.) Barthes's insights accord with the intuitional logic of our approach. Rather than emphasising opposition (or binary-opposition) with the choice of contrasting colours, a linguistic map should use a thoughtfully chosen colour-palette which serves to reinforce the range of comparable or equivalent semantic connections to be made between the data. Thus, instead of foregrounding polarity in language, or the hierarchy implied by a red-blue contrast, we contend that a more carefully balanced, non-contrastive palette better facilitates an understanding of the complex relations in a language continuum.

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." For instance, as already mentioned above, the name of the Flanders Moss National Reserve' might serve as an explanation for the occurrence of 'skelf', an item with Dutch roots, in the respective area (cf. [Hessle, 2019: 13]). 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 and 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.

III. PRESUPPOSITIONS ABOUT 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 the restrictions of the questionnaire already imposed 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 and 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 that of the LAS 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 and 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). While such descriptive 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.

3.2. 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. Figure 7 shows the biodata of the original respondents; Figure 8 shows a scanned version before editing. 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 7 and 8). 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 process of digitisation must

involve the identification of potential irregularities and the implementation of a quality control regime, for example in form of a manual one-by-one comparison of potentially disrupted digital material with its analogue original, in order to ensure the integrity of the language data. The data-set provided for automatic machine reading must follow an extremely rigid structure, since, as the above instances show, even 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	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	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 7. Data-list from *The Linguistic Atlas of Scotland* [Mather and Speitel, 1975: 400]

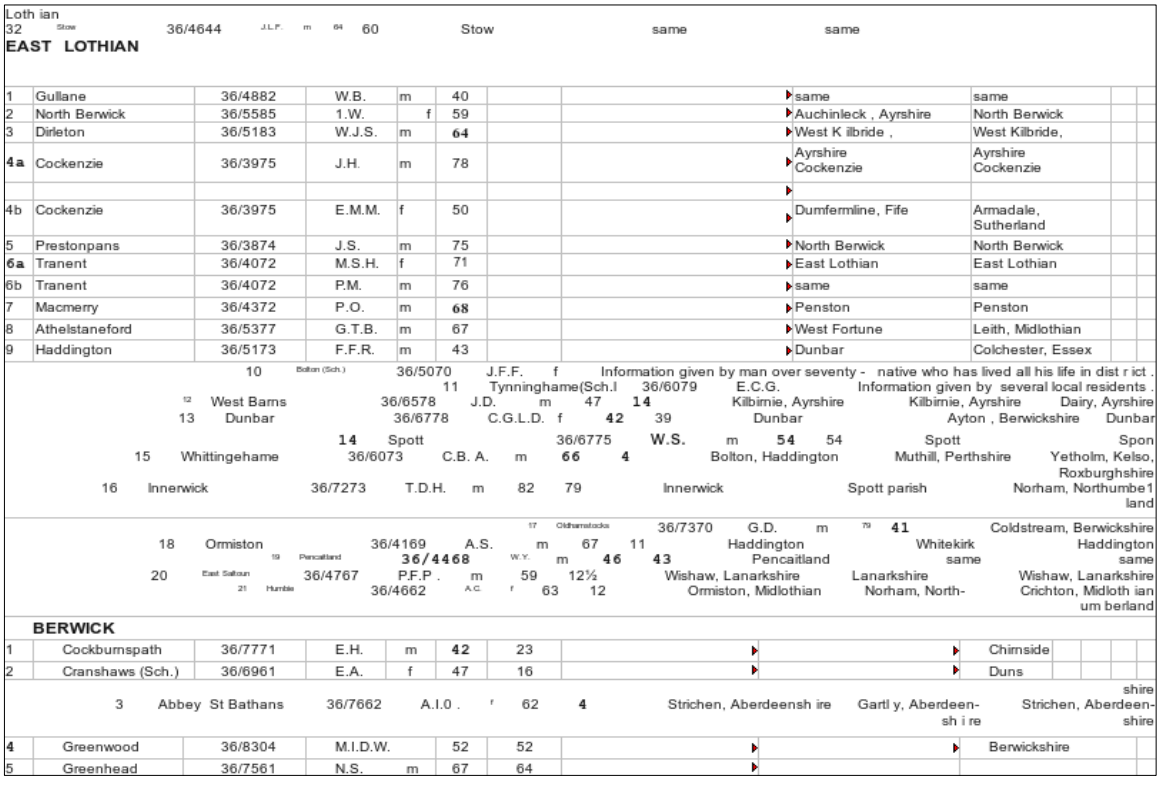

Figure 8: OCR-output from *The Linguistic Atlas of Scotland* data

Conclusion

This study shows that it is possible and necessary to establish a theoretical framework for the digitalisation of a historical data collection. For creating an atlas, our model distinguishes between topological issues (such as those pertaining to the data) and topographical issues (such as those pertaining to geography and the physical terrain). We further distinguish between the physical map and the cartography being superimposed on it. We contend that it will be from interpreting the relationship between the cartography and the physical map – i.e. that between the topology of the data and topography of Scotland – that insights into the nature and distribution of Scots will emerge. We anticipate that the fresh insights thereby enabled will go some way towards providing the kind of lexicological study of Scots on the basis of the atlas material as was envisaged at the outset by the original compilers.

We have also indicated some of the advantages of a digital atlas, such as the possibility to visually contextualise the linguistic data by combining it with extralinguistic information, enabling the derivation of yet more new results. However, we stress that all results necessarily reflect back on the categorisation and systematisation of the data-set, on the basis of which any and all interpretation will inevitably be made. We show that, in the course of the digitisation, processes whose nature is often perceived as purely technical must necessarily be questioned, including, for instance, the choice of a colour-palette or the checking and correction of data input. Technical care is even more so required, given the narrow limits of digitisation technologies, despite their manifold advantages.

ACKNOWLEDGEMENT

We are grateful to the permission given to John Kirk by Croom Helm to digitise the atlas maps and data. We are indebted to the editors for their invitation to contribute to this Special Issue, and to two anonymous reviewers for their stimulating criticisms and many helpful suggestions.

References

Allred, S.C., Schreiner, W.J. Smithies, O. Colour blindness. Still too many red-green figures. *Nature* 2014;510:340.

Amadae, S. M. *Prisoners of Reason. Game Theory and Neoliberal Political Economy*. Cambridge University Press (Cambridge), 2015.

Bejinariu, S.-J. and Olariu, F.-T. Romanian Linguistic Atlases in Digital Format – A New Approach. *Philologica Jassyensia* 2017;XIII(1/25):13-23.

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

Badiou, A. and Haéri, G. *In Praise of Mathematics*. Polity Press(Cambridge), 2016.

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

Concise Scots Dictionary. (2nd edition). Edinburgh University Press(Edinburgh), 2017.

Dictionary of the Older Scottish Tongue. Electronic Version. http://www.dsl.ac.uk [20 Dec. 2018].

Encyclopædia Britannica. Luitzen Egbertus Jan Brouwer. 2019. https://www.britannica.com/biography/Luitzen-Egbertus-Jan-Brouwer [20 Feb. 2020].

Flegg, H.G. *From Geometry to Topology*. The English University Press(London), 1974.

GeoJSON. The GeoJSON Format. 2019. https://tools.ietf.org/html/rfc7946 [19 Jan. 2019].

Gorgescu, G. N-Valued Logics And Łukasiewicz–Moisil Algebras. *Axiomathes*. 2006;16:123-136.

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

Hill, A. Game of Bones: Thousands descend on rural Saskatchewan town to throw horse ankle bones. *Saskatoon StarPhoenix Electronic Edition*, 2 Aug. 2016. https://thestarphoenix.com/news/local-news/game-of-bones-thousands-descend-on-rural-saskatchewan-town-to-throw-horse-ankle-bones [20 Dec. 2018].

Hoffmann, Dirk W. *Die Gödel'schen Unvollständigkeitssätze. Eine geführte Reise durch Kurt Gödels historischen Beweis*. (2nd edition). Springer Spektrum(Karlsruhe). 2017 [2013].

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

Kirk, J.M. East Central Scots: A Computerised Mapping Package. Fenton, A.S. and MacDonald, D.A. (eds.). *Studies in Gaelic and Scots.* Canongate Academic(Edinburgh), 1994a:48–68.
Kirk, J.M. Maps: Dialect and Language. In Asher, R.E. Asher and Simpson, J.M.Y. (eds.). *The Encyclopedia of Language and Linguistics. Volume 5*. Pergamon Press(Oxford), 1994b:2363–2377.
Kirk, J.M. 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 2019.
Kretzschmar, W.A. Isoglosses and Predictive Modelling. *American Speech* 1992;67(3): 227-249.
Leaflet. 2019. Leaflet – a JavaScript library for interactive maps. https://leafletjs.com/index.html [19 Jan. 2019].
Leck, I. Is a minor inconvenience. *British Medical Journal* 1994;309:129.
Leonard, R. *Von Neumann, Morgenstern, and the Creation of Game Theory. From Chess to Social Science, 1900–1960*. 2010. Cambridge University Press(Cambridge).
Little, W. and Onions, C.T. (eds.). *The Shorter Oxford English Dictionary on historical principles.* (3rd edition). Clarendon Press(Oxford), 1956.
Macaulay, R.K.S. Review of Mather, J.Y. and Speitel, H.-H. (eds.). *The Linguistic Atlas of Scotland: Scots Section. Volume 1.* Croom Helm(London), 1975. *Language*, 1977;53:224-228.
Macaulay, R.K.S. Review of Mather, J.Y. and Speitel, H.-H. (eds.). *The Linguistic Atlas of Scotland: Scots Section. Volume 2.* Croom Helm(London), 1977. *Language*, 1979;55:224-228.
Macaulay, R.K.S. Linguistic Maps: Visual Aid or Abstract Art? In Kirk, J. M., Sanderson, S. and Widdowson, J.D.A. (eds.). *Studies in Linguistic Geography: The Dialects of English in Britain and Ireland*. Croom Helm (London), 1985:172-186.
Mapbox. 2018. About Mapbox. https://www.mapbox.com/about/ [19 Jan. 2019].
Mather, J.Y and Speitel, H.-H. (eds.). *The Linguistic Atlas of Scotland. Scots Section. Volume 1.* Croom Helm (London), 1975.
Mather, J.Y and Speitel, H.-H. (eds.). *The Linguistic Atlas of Scotland. Scots Section. Volume 2.* Croom Helm (London), 1977.
McClure, J.D. The Linguistic Atlas of Scotland. *American Speech* 1976;51(3-4):223-234.
Murison, D.D. 1978. Review of Mather, J.Y. and Speitel, H.-H. (eds.). *The Linguistic Atlas of Scotland: Scots Section. Volume 2.* Croom Helm(London), 1977. *The Scottish Review,* 1977;9:45-48.
OpenStreetMap. History of OpenStreetMap. 2018. 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, M. Crowsteps in Fife: The Flemish Connection, Part 1. *Scotland and the Flemish People Project.* St. Andrews: St. Andrews University. 2013. https://flemish.wp.st-andrews.ac.uk/2013/12/06/crowsteps-in-fife-the-flemish-connection-part-1/ [10 Jan 2019].
Scottish National Dictionary. Electronic Version. http://www.dsl.ac.uk [20 Dec. 2018].
Upton, C. and Widdowson, J.D.A. *Lexical Erosion in English Regional Dialects*. National Centre for English Cultural Tradition(Sheffield). 1999.
Xie, Y., Aristar-Dry, H., Aristar, A., Lockwood, H., Thompson, J., Parker, D. and Cool, B. 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*. 2009:305-311. [downloadable from https://fedcsis.org/archive/proceedings_2009]