NATIONAL WORD INVENTORY: NOTES FOR A HISTORY

Matthew Fuller

The National Word Inventory (NWI) data set consists of word samples and accompanying standard observations obtained by word surveyors in Great Britain between 1978 and 1982, sampled at a 5-mile interval, based on the Ordnance Survey National Grid coordinate system. In retrospect, the work of the NWI seems to anticipate many current debates that surround big data, statistical models and environmental monitoring, and is the subject of renewed interest. In undertaking the work, surveyors would attend the location at a precise grid reference point and then note the first word overheard by them. They were instructed to record only the first word that they accurately heard. Researchers were local people selected after rigorous auditory competence tests. They were drawn from the local population to ensure their attunement to any variation in accent or dialect. When attending the specified location and carrying out the audit, the researchers would complete a standard one-page form on their clipboard, and send it in a pre-paid envelope to the central NWI office in Wisbech, where an archive was kept, and analysis was carried out.

The NWI survey was highly successful in attaching specific linguistic units to particular geographic locations. Undertaken on a monthly basis, sixty times at each point, this survey is the first systematic analysis of the correlation of place to language, anywhere in the world. The survey is often said to show the relative equilibrium of vocabulary throughout the survey period. Much of the word use is statistically aligned with the mean usage of the words recorded in other sources of that era, for instance via the analysis of recorded oral history or word-frequency analyses of novels, newspapers and other written sources in the public domain. Nevertheless, there are some remarkable findings. By analysing this data set we can see that the word 'always' slowly moves across the British landscape in what may be seen as an inclined wave form throughout the whole of 1978. It then disappears from usage, as recorded in the data set. The term "cough syrup," understood as a euphemism–the meaning of which is not given by the recorder, owing to observation of the experimental protocol, which imposed strict limits on interpretation–appears

twice, in an alley to the rear of Aboyne Place in Aberdeen, and in the nave of Coventry Cathedral. Other notable data shows, inter alia, that an observation point on the westward side of Brynhafod Playing Field in Oswestry displayed the highest use of the word "curtilage" in the NWI; the location situated at the northward end of the beach at Holkham, in Norfolk, alternated twelve times between records of people saying "water" and "sky" before the pattern dissipated, to be replaced by "each" for the next five samples, before the remaining samples exhibited statistically average mixes; urban areas, taken as a whole, tended to favour shorter words, whereas those in the north-east of England exhibited the largest use of consonants per word of any area. No meaning has yet been attributed to this.

From the sample facts noted above, one can readily see that this data set is exceptionally rich, and lends itself to multiple methodologies and modes of enquiry in linguistics, geography, communications studies, sociometrics and information science. Researchers in all these fields made notable findings through their use of the NWI data set during these "golden years."

Researchers assigned to monitor the more remote parts of these islands would sometimes situate themselves for extended periods of time before a word made itself available for auditory inspection and textual recording. In the data gathered from these locations, there is some evidence to support the presence of a small number of typical experimental artefacts, such as researchers mistaking wind-ear interactions for spoken words, or researchers spontaneously uttering a word upon reaching the maximum time allotted in position (24hours) without otherwise having overheard any data, and being forced to self-record. The latter kind of utterance was sometimes understood to be a source of shame to those concerned, but was looked upon in a kindly fashion by their colleagues in busier, and hence more voluble areas. In contrast, the phenomenon known as "The Tewkesbury Absence" denotes the seeming lack of speech at any location within a five-mile radius of the epicentre of this thriving and well-loved Gloucestershire town throughout the entire period of the NWI's field research. Attempts to "back-fill" the data using various forms of statistical regression suggest that it is difficult to know, with any satisfactory degree of accuracy, what was being said in this otherwise quantitatively unremarkable zone during this period.

After a fallow period in geo-spatial vocabulary research following the closure of the NWI on the 1st of January 1983, it is understood

that enthusiasts, many of them NWI veterans, are planning to breathe new life into the field by digitising the archive and making it available online as an augmented-reality test case. Users will be able to travel to the locations in which the words were originally recorded, and listen to their enunciation by a computer-synthesised voice as many times as they please. They may also choose to make up "sentences" by moving from one place to another, across the grid reference system. The latter approach would make the spatial availability of articles such as "the" and "it," and of conjunctions such as "and" or "or" of particular significance when choosing routes. Happily for such purposes, recordings show that these words are rather evenly distributed across the territory. A further new and exciting phase of this research that uses contemporary tools is being inaugurated by researchers who use machine learning to calculate the probable words being spoken by people–whose faces are blurred to public view–whose mouth-postures are visually recorded by Google StreetView. The sometimes-lively public response to the data-gathering vehicles makes this an especially rewarding data set to peruse and to analyse.