The validity of the Linguistic Fingerprint in forensic investigation

Rebecca Crankshaw (Linguistics)

Abstract

This article is concerned with the definition and validity of a linguistic fingerprint, a relatively new notion that remains both undeveloped and undefined. It is argued that interpretation in the literal sense is misleading, as it seems to imply that wholesome samples of a person’s language use can be collected and that an individual’s language use is fixed (Coulthard 2004). Others (Chomsky 1965, Halliday 1975, Anshen 1978, McMenamin 2002) support the notion’s existence, drawing on the importance of sociolinguistic variables, as well as an individual’s preference of selection. These factors are accepted by Coulthard (2004) although merely as a guide, or an aid for selection/de-selection of an individual rather than validating the existence of a unique identity marker. All this considered, the article concludes that, taken in the literal sense, the notion lacks any possible validity and, although the idea of it being used as an aid of identifying an individual is perfectly viable, this in actual fact does not exceed the current practices already in place and is therefore not a new notion at all.

Keywords: Linguistic Fingerprint, Forensic Linguistics, Applied Linguistics, Authorship Identification, Idiolect.

Forensic Linguistics is a relatively new area of study within Modern Applied Linguistics, using scientific linguistic study in forensic contexts (McMenamin 2002). Due to its short history, new and continuing research has meant that the area will continue to grow and develop and inevitably so will the classification of areas within the field of Forensic Linguistics itself. Its applications currently include areas such as voice identification, the language of the law and analysis of discourse in legal settings, drawing together the techniques derived from syntax, semantics and pragmatics (concerned with language structures, meaning of words and the usage of such words respectively), phonetics and phonology (relating to the formation of sounds used in language and how they create meaning), and corpus linguistics (corpus linguistics is the study of language through the use of real life examples).

The lack of extensive research alongside its short existence means that the extent of its possibilities and limitation cannot be fully known. Therefore, whilst Forensic Linguistics is capable of being used as evidence in court, it is usually a much more successful defence and cannot be used as the sole piece of
evidence for prosecution. There are areas within Forensic Linguistics in which much research is yet to be done and subsequently new notions and theories to arise. One of these areas is within Authorship Identification and the idea that it may be possible for the existence of a Linguistic Fingerprint, which will be the focus of the following content.

The notion of a Linguistic Fingerprint is a very new and undeveloped concept, based on the idea that each individual has a unique way of speaking and writing by which they are identifiable. As yet, there appears to be no concrete definition for what amounts to a Linguistic Fingerprint and this very fact contributes greatly to whether such a notion can be deemed viable or not. However, if we take as the definition ‘an identifiable unique or distinctive way in which each individual speaks and writes’, then support for the theory may be viable.

As there are variations of linguistic forms between dialects, it seems sensible to conclude that if such variation in a language exists at group level, then similar variation should also be found at individual level. This is a position supported by Anshen, stating that: ‘Not only do any two members of what should reasonably be the same speech community use different variation of the same linguistic form, but so does each individual member’ (Anshen 1978, 1). It is for this reason that ‘the idiolect has been referred to as the personal dialect’ (McMenamin 2002, 56).

McMenamin defines the idiolect as being ‘the individual’s unconscious and unique combination of linguistic knowledge, cognitive associations and extra linguistic influences’ (2002, 53). This shows acknowledgment that external factors affect the way that a person uses language whether it be through geographical location, social class, level of education, age, sex and so forth, all subjects of Sociolinguistic studies which have shown that such external factors do contribute to the language that a speaker/writer uses. As it is expected that no two individuals will have had inputs of exactly the same external factors, it is also expected that they will not use and perceive language in the exact same way as one another. Instead, at the very least, it is believed that there will be small or subtle differences in their internalised grammar which then manifests itself in a person’s speech, writing and responses to others (McMenamin 2002, 53).

The above shows how variation exists at the very core of language, but it also exists at other levels of language as well as the underlying structures. This can be seen in word selection, meaning the words
one chooses to use and when to use them, also a result of contributing external factors. This is because language learning is constant and over time an individual builds up something referred to as their active vocabulary, which is very large and effectively contains an extensive list of words that have been encountered and learned by the individual. Due to the vocabulary being something that is dependent on external influences, it is a reasonable assumption that each person’s active vocabulary will differ at some level, whether it be through the actual items that each individual has stored and/or their preferences in the selection of such items instead of others (Coulthard 2004, 1).

From the aforementioned it is clear to see that individuation does exist in language and Sociolinguistics offers much support for this. It could be said that Sociolinguistics formed part of the birth of the concept of the Linguistic Fingerprint, as it has drawn us towards the conclusion that each individual possesses their own idiolect (a person’s own variety of language), which Halliday et al (1964) and Abercrombie (1969) suggest will manifest themselves through ‘distinctive and idiosyncratic choices in texts’ (cited in Coulthard 2004, 1), and so in theory each speaker/writer should be uniquely identifiable.

However, uniqueness in writing has been documented to be not only something that is found from one individual to the next, but also within an individual themselves. Chomsky (1965) and Halliday (1975) refer to this as the ‘uniqueness of utterance principle’, which suggests that not only will two individuals writing about the same topic produce texts which are unique from one another, but so will the same individual on two separate occasions, as they are likely to make a different set of lexico-grammatical choices (Coulthard 2004, 3).

This adds complication to the notion of a Linguistic Fingerprint, as it may be just as likely to conclude that two texts written by the same author are indeed different writers and the idea of an individual idiolect may not be as powerful and dynamic as first seemed. Even within an individual their language use is subject to change, as they are constantly subjected to external influencing factors that have a direct effect on language. For example, should it be expected that an adolescent having not yet completed their education will use the exact same grammatical and lexical choices ten years later? Similarly, it is known that variation exists between dialects in terms of both grammatical structures and lexical items. With this in mind, is it a reasonable expectation that a person who moves from one geographical location will to not add any dialectal traits from the later location to their current active vocabulary?
This is why the definition of a Linguistic Fingerprint is fundamental in concluding whether or not it exits. This is because of the presupposition of the word ‘fingerprint’ and the inevitable comparison to the physical fingerprint. The physical fingerprint renders a person uniquely identifiable, as each sample is both ‘identical and exhaustive’, containing all of the information necessary for an individual’s identification. In contrast to this, even an extremely large sample of linguistic data can only ever give partial information of a person’s idiolect (Coulthard 2004, 2).

Coulthard addresses this presupposition of ‘fingerprint’ by dubbing it a ‘misleading metaphor’ and stating that it leads us to the misconception of huge databanks being created that contain representative linguistic samples belonging to millions of idiolects, which like the principle of a physical fingerprint in a database, questioned texts can be tested and matched against in order to confirm authorship (Coulthard 2004, 2). In reality, to collect such a wholesome linguistic sample which would represent a person’s idiolect in its entirety would not at all be possible. Unlike the physical fingerprint which is a complete and fixed identity marker from birth, a person’s vocabulary is constantly growing and changing whilst grammatical structures differ in complexity throughout an individual’s lifetime. It would be impossible to collect linguistic samples to encompass every possible aspect of a person’s language use in every situation that they may encounter. On these grounds, and with the definition of a Linguistic Fingerprint to be taken in the literal sense of a unique and personal identity marker, then the concept seems to be flawed.

Aside from what the term ‘Linguistic Fingerprint’ infers, the notion raises multiple questions. In terms of what it is defined as, if it is proposed to be a verbal equivalent of the physical fingerprint then the evidence opposing the theory currently seems to outweigh the supporting evidence. If it is meant in terms of a broader sense of language uniqueness that gives clues towards the identification of the author, then although it would be a more viable theory, does it actually amount to a new theory? In practice it would not demonstrate anything more advanced than what is already in existence. This is because currently a Forensic Linguist’s task, as Coulthard (1992, 1993, 1994a, b, 1995, 1997) and Eagleson (1994) point out, involves selecting and even deselecting an author from a small number of possible candidates based upon supporting linguistic evidence which may confirm or deny attribution to one of them. It is never a case of asking ‘of all the millions of speakers of English, which one produced
this particular text?’ (Coulthard 1998, 141) and it seems clear that this interpretation of the theory offers nothing more than the current system already in place.

Furthermore, regardless of its questioned validity, due to the concept being in its infancy, there is not yet so much as speculation as to what kind of data would need to be collected and analysed in order to create a unique categorisation of an idiolect (Biber 1988 & 1995, Stubbs 1996 cited in Coulthard 2004, 2).

Similarly, there have been limited suggestions as to the quantity of text needed in a questioned piece of writing in order to exhaustively specify an individual’s idiolect in a successful manner. Authorship investigation cases usually tend to be centred around a text which is often very short in length and although Morton (1986 & 1991) asserted that the ‘fingerprint’ was something that could be derived from extremely short pieces of texts claiming that it was ‘inherent in the basic blocks of text’ (cited in Coulthard 1998, 141), his methodology was later found to be flawed. Now, the current understanding is that a substantial amount of text is needed in order for the analysis of their lexical items and collocations to be possible. As Forensic Linguists rarely received texts longer than 750 words, their traditional analyses are not always possible as the relevant items are a rarity in such texts and so alternative analysis must be used (Coulthard 1998, 142). Whilst the text can still be analysed in terms of other items, the fact that some analysis cannot take place does impose limitations on how an individual’s idiolect is represented, certainly prohibiting an exhaustive representation from being possible.

In light of the evidence both in support of and opposing the concept of a Linguistic Fingerprint, it is apparent that this is an area in which more research is needed. Based on the research that exists to date, it is also apparent that the lack of a definition makes the validity of the concept hard to judge. If it is to be taken as a unique identity marker by which to pin point the identity of a writer then it seems both impractical and impossible. If it is simply proposing that identity markers can be found within texts that can then act as an aid to selecting or deselecting an individual as the author, then it does not exceed the current mechanisms for solving authorship disputes. In any event, regardless of how it is defined, the concept as it stands does not appear to be able to offer anything further to the field of Forensic Linguistics with it being deemed ‘an unhelpful concept’ and a ‘misleading metaphor’ alongside
a claim that for the disputed authorship cases which currently exists, ‘one does not need anything as sophisticated as a linguistic fingerprint’ (Coulthard 1998, 141)

Additionally, there are multiple implications and limitations imposed on the concept brought about by the fact that the types of data needed are unknown, and further problems are added when how much data would be needed is given contemplation. It would be incomprehensible to collect data in such a way that was able to represent the entirety of a person’s idiolect, revealing further flaws surrounding the notion. Even if this was thought to be possible, the extreme lack of disputed texts of an adequate length for an analysis of such depths would mean that an accurate comparison of questioned material against the representation of idiolect would not have the required level of accuracy to prove authorship undoubtedly.

With all this considered, it seems that the amount of uncertainty and limitations surrounding the notion have undermined the concept to such a level that – at least for now – it cannot be considered viable.

References


Coulthard, M. 1998. ‘Identifying the Author’, *Cahiers de Linguistique Française* 20, 139-161.


