The Role of Linguistics in Veracity Evaluation

Martina Nicklaus and *Dieter Stein**

Abstract

Although the issue of veracity of statements in forensic contexts arises from evidence given in language, an appreciable share of applied professional linguistics in the practical process of veracity evaluation hardly exists. The article aims to provide a survey of key theoretical and methodological issues in the field of veracity evaluation from the linguistic point of view – a currently popular field where fast results are expected and a lot of publicity is to be gained. The article starts out by looking at the very notion of "truth" and what can happen to it on the way from real-life experience to a verbal report. It then gives an overview of psychological and other approaches to verifying the truth or otherwise of verbal reports. These approaches variously include linguistic cues in their lists of diagnostic features. Dissatisfaction with and inconsistency of the results of psychological tests is partly due to a lack of professional sophistication in defining linguistic cues, as well as a disregard for the role of genres in determining normality expectations for the occurrence of linguistic cues. The paper argues for the inclusion of linguistic cues on a systematic basis, as well as for the further refinement and sophistication in the definition and application of linguistic categories in psychological tools of analysis.

* *Martina Nicklaus*: Heinrich-Heine-University Düsseldorf, nicklaus@phil.uni-duesseldorf.de; *Dieter Stein*: Heinrich-Heine-University Düsseldorf, stein@phil.hhu.de.

1. Purpose

The present paper discusses the role of linguistics in veracity evaluation. The field of veracity evaluation is in practice dominated by a number of psychological procedures. We argue that adding technical linguistic knowledge to the toolbox of veracity evaluation is necessary to analyze evidence given in the very medium that is the object of scientific study of language. The issue of veracity of statements given in language is the more at the center of interest where this evidence is the only or nearly only evidence, such as often in cases of child abuse.

Linguistic expert knowledge seems essential from two points of view. Some of the techniques employed in psychological analyses use linguistic categories (e.g. correction strategies or the expression of emotion). We will argue that a refinement and further differentiation of these categories from the linguistic point of view will yield better results in psychological analysis, including enhancement of possibilities for automatic analysis. In addition, a range of possible linguistic categories of analysis seems to be completely unexploited for this type of forensic analysis.

A word of caution is necessary. It will never be the case that there will be a one-to-one categorical relationship between veracity and a given linguistic form (say "pronouns" or "passives") enabling some sort of blanket "algorithmic" analysis. The reason is that the occurrence of linguistic forms has always to be gauged against genre baselines. A grammatical passive, for instance, is to be expected in some types of discourse, and its occurrence bears no special information there, while in others it is an information-bearing deviation from a text norm. So an intimate knowledge of baselines, ideally manifested of specialized corpora, is a prerequisite for the application of technical linguistic knowledge and fruitful complementation of psychological with linguistic knowledge. Additionally, the assessments should take into account the personal baseline of the presumed liar, that is his verbalizing preferences when telling the truth.

This paper will first give a rough overview of some basic concepts relating to "veracity" and of the state of the art in psychological analysis, then proceed to discuss linguistic issues as they arise in the application of psychological analysis, and finally use two examples from linguistic analysis to show how a refined application of linguistic knowledge is likely to yield more satisfactory results.

Another aspect has to borne in mind. The issues as they are discussed here are such as they appear from the analysis of Western, especially Anglophone and Germanophone cultures, and as such have an inherent cultural bias. This relativity must be stressed in a political and legal situation where forensic linguistics must adopt a more global view of issues, including issues arising from migratory processes. This relativity of our discussion applies obviously to issues of the structure, the meaning and use conditions of linguistic forms, but it also applies to the cultural basis of the psychological and emotional categories, such as the expression of emotions in the context of deception (Taylor et al., 2017: 3).

2. On Truth

At the center of evaluations of veracity is the "truth". However, the truth about "truth" is not a simple matter, for several reasons. Apart from the linguistically or pragmatically inherent problems to be discussed in the following, notions "truth" are also relative to the type of context: what the naked facts are (if they are indeed postulated to exist as such), what it is in the perceptions of participating individuals and what it is for the court. In addition, renderings of "true facts" are also relative to the type of linguistic genre in which they are embedded: if in a cross exam or in one of the many subtypes of narrative genres – to the friend, to the psychologist, if rendered once or several times in different contexts. Also it makes a difference whether the "truth" is one told by a child or an adult. All these subdifferentiations would deserve separate dedicated treatment in separate chapters or books. Given the limited space of the present discussion, our primary focus will be limited to the methodological issues, and further differentiations into specific situations such as above will only be referred to if obvious and undisputed. Apart from the cultural relativity of notions of truth and the concept of a "true story" further differentiations are in place from the linguistic, especially the pragmatic, point of view. As straightforward as it seems at first go, and arguably to most practitioners, the simplest phrasing of the issue – *"Does a witness lie or is she telling the truth?"* – is problematic. It is not as easy as one might think to define a lie from the side of an adult (Horn, 2017). In addition to the many different types of lying discussed by Horn that involve the actual use of language, profanely speaking, "actually "saying" something that can be true or false, you can also lie by *not* saying something, i.e. by not producing a locutionary act. Lying by *not* offering relevant information can distort the "story", the judicial narrative, just as much as producing statements that are factually false. The key notion here is "relevant": it involves a shared interest of speaker and hearer. In Gricean terms, it involves adherence to the maxim of relevance. Situations like interviews with the prosecutors and at court are characterized by very specific constraints on relevance, and the presumption of mutual sharedness or not.

Just as any notion of "baseline" with respect to privileges of occurrence of linguistic forms can never exist *in abstracto*, but is very relative to trans-individual and societally shared notions of genres, so the notion of relevance dependent on notions of what is considered relevant information in different types of contexts in the legal process. In the case of deliberate lying there may be a mutual awareness of a fight about what is relevant information.

But the notion of lying in the case of a child, say, a 14 year old, or younger, has added complexities due to several types of developmental factors, ranging from the ability to empathize, i.e. to gage the other person's mind, without which you cannot be intending to deliberately lie, to the availability of linguistic structures that may be involved in lying.

For an adult, a lie – as separate from deception – is best defineable as uttering a proposition that the speaker is committed to the un-truth of. The moral default case is

uttering propositions the speaker is committed to the truth of. So, for the linguist, you lie even if what you say happens accidentally to be true. Accidental truth is more difficult to handle for the legal world. Among linguistic pragmaticians there is agreement that what matters is the intention to lie. What matters is that you are committed to the un-truth of the proposition, in the case of a lie. The adult does this with language and with the intent to state something that is not true.

You may not lie, but you may still not tell the truth. The legal process has two basic types of interests in veracity evaluation: the judge wants to know, not, negatively, what is not true, but, positively, what facts ("rechtsrelevante Fakten", i.e. norm-defining facts to be considered relevant to decide the case) can be used to base the judicial narrative on. The other, related, judicial interest is in identifying and punishing perjury. While the prototypical intentional lie is, apart from its potential deleterious effect on the judicial narrative, once identified, a case of perjury, there are a number of processes and factors that may cause departures from what was the original "fact", which may be "untruths", but are not lies in any intentional sense. Intentional lies may be perpetrated by children from a certain age onwards, with adolescents and adults as "mature" liars, but there are a number of processes in the generation of a witness report or a narrative that cannot be attributed to intentional lying and are consequently not candidates for perjury. While these processes are observed in both adults and children, some are more typical for children and adolescents, like "false memories" (discussed in 3.2.).

In addition, and in partial overlap between psychological and linguistic perspectives, there are a number of approaches to the origin of stories that linguistic research has unearthed and which veracity evaluation will ignore at the cost of validity. If, as veracity evaluation aims to do, we want to retrace from the "surface" of a linguistic account the factual history we need to be aware of the "route" of an initial experience to a linguistic report at a later time stage. Quasthoff (1980) has mapped out different processing stages in which a factual aspect of experienced history is first interpreted and stored in a "cognitive history" and in later recalls summoned up and reproduced in different production and utterance contexts. While psychologists are entirely aware of how a slice of personal history is "experienced" and interpreted, pragmatics has emphasized to what extent the rendering of past experience (of content of cognitive history) is dependent on the properties (social, psychosocial, emotional) of the external situation and utterance purpose.

In a similar vein, Eades (2012) has demonstrated to what extent the production of any utterance and certainly of a verbalized history is the result of "co-creation" of the discourse. She demonstrates that in the end the reported "facts" can be significantly different from what was earlier on reported and what were the independently established "facts"; Eades' conclusions correspond to the findings in the studies of Hannken-Illjes (2015) and Günthner (2005). She can base her discussion on a body of empirical research such as Schiffrin (2006) and Norrick (1998) that it is "a naïve assumption" that "people always tell the same story in the same way" and "that no two tellings are the same: there

are shifts in perspective and different details are included or omitted, highlighted or backgrounded" (Eades, 2012: 476).

A related finding from another type of discourse casts doubt on a criterion employed for the veracity of verbal evidence. The variations in the elicitations of a report were taken as indications of falseness or lying. The difference between two versions of the same statement offered at different times ("They [the police] grabbed the three of us" and "they were told to get in the police cars", Eades, 2012: 481) makes a great difference in the credibility assessment of the witness and being accused of lying or not.

In the light of the findings of Eades (2012) it is clear that this argumentation is not reliable. For instance, during the whole judicial process the witness had to repeat the account five times in different legal contexts (Eades, 2012: 381–384). In other words, five different accounts had to be co-created with different contexts of usage. Given the processes that intervene between the formation of the cognitive history and each verbalized reproduction, which in themselves can effect changes in the cognitive history, we cannot a priori assume that there is a stable and immutable content, or even a "text", that simply needs digging out on successive occasion.

This is also why decontextualized, experimental, methods of pilot-testing of quantitative procedures cannot yield any valid results. While German SVA experts welcome more empirical studies in principle (Steller & Köhnken, 1989: 241), they deny any value of laboratory studies that evaluate single criteria (Undeutsch, 1984: 63; Greuel, 2001: 320). This is immediately understandable, given also the special psychological situation of abuse victims, which never could be simulated in experiments, and be it only for ethical reasons. A verbalized story is a fully pragmatically specified process, not even an object ("a text") that will allow the identification of factors as a physical object that can be controlled for all kinds of conditioning factors. All quantified tests generally assume to some extent the existence of such a robust and immutable content layer that only needs to be tapped by investigative methods and that is part of a broader, culture-dependent ideology of a "true story" (Eades, 2012; Trinch, 2003):

> "Trinch's work on lawyers and paralegals producing written statements on the basis of interviews with clients establishes the "ideology of narrator authorship" ([Trinch] 2003: 49–50): the assumption that a witness' or interviewee's story is solely their own account. She points out that it is found not just in the culture of the law, but more generally in Western culture. This ideology relies on the "prevalent" and "tenacious" cultural notion of the "true story", but ignores the collaborative nature of storytelling." (Eades, 2012: 478)

So it is for very good reasons inherent in the nature of stories as representing passage from experience to verbalization that truth evaluation has shifted from categorical predicates to terms like "experience-based", in addition to psychologically well-established phenomena like "false memories", to be further discussed in section 3.2.

3. Statement Analysis

Forensic theory and practice are dominated by psychological evaluation in different versions of the discipline. The prevailing method in germanophone countries is *Aussagepsychologie* ("psychology of testimony", Undeutsch, 1984: 51) and, more precisely, its main method *Statement validity assessment* (SVA, "Glaubhaftigkeitsbegutachtung"[1]), an authorized technique for experts' reports at court. In anglophone countries some roughly comparable methods for veracity assessment are used for police interviews: *Scientific content analysis* (SCAN) or *Reality Monitoring* (RM); recently the *cognitive lie detection approach* (Vrij, Fisher & Blank, 2017) has been proposed as alternative. The probably latest study in the field of deception detection, published in 2018 by Quijano-Sánchez et al., assesses the validity of a tool, *VeriPol,* that is specially designed to detect insurance fraud (see also below, section 5).[2]

The users of these tools implicitly define lying (SVA, *VeriPol*), or deception (RM[3], SCAN[4]) as methods to identify intentional statements of untruths (Meibauer, 2014: 106), or tactics to intentionally hide the truth (Meibauer, 2014: 107–154). In what follows we adopt this operational definition for the purposes of the further discussion, bearing in mind the problems of this definition in a larger context of discussion as in section 2 above.

The above-mentioned tools used for deceit detecting comprise a cue-based analysis of the verbal part of statements, in part involving the analysis of linguistic structures. In SVA, this linguistic part of the analysis is, by common consent, carried out with good success in court practice. However, there remain substantial questions and uncertainties with respect to the applied linguistic veracity cues, as several meta-studies or reviews confirm (Welle et al., 2016; Hauch et al., 2017; Vrij, 2005; 2009; 2014). Not without very good reason German SVA experts, whose reports are highly valued at court and often decide cases[5], are extremely careful before presenting any assessment. They apply the list of cues, the *Content Criteria,* only to transcripts of "considerable length" (Hauch et al., 2016: 820; similar: Greuel, 2001: 213; Vrij, 2005: 15), that is of no less than 15 pages, as far as we can judge from our interviews with psychological experts. Shorter testimonies are not even taken into account and are discarded as evidence.

---

[1] Fitzpatrick, Bachenko & Fornaciari (2015: 32) translate as: "Statement validity analysis".

[2] The authors tested their tool *VeriPol* on no less than 1,000 reports presented to the Spanish police in 2015. VeriPol includes the scanning of word combinations (unlike LIWC) and performed better than current tools or human evaluators, identifying false reports in about 90 %. This promising result however might not be reproducible for genres that are more complex and uncalculable with regard to content and involvement, as oral reports on abuse.

[3] Originally this method was developed to distinguish "memories for thoughts and memories for perceptions" see Johnson & Raye, 1981: 67.

[4] A method supposed to discover deception in police interviews, see the homesite of SCAN, lsiscan.com/id37.htm.

[5] Steller & Köhnken (1989: 235), report that in 90 % of the by then known cases the judge had followed the expert's evaluation.

Statement analysis is conducted to evaluate the veracity of oral, written or transcribed statements made for the police, for psychological experts or at court. The various techniques for statement analysis (see the comprehensive overview in Vrij, 2009: ch. 7–13) all rely on particular linguistic, behavioural or merely physiological cues as indicators of veracity – or of lying and deceiving. In the following, we focus on the technique used in germanophone countries, pointing, where appropriate, to the aspects shared by competing models.

3.1. Content Criteria

Aussagepsychologie ('psychology of statement/testimony') is a branch of forensic psychology that developed in Germany and Sweden at the beginning of the 20th century (Undeutsch, 1967: 29) to unite issues dealing with veracity assessment. One of its most common techniques, originally and until today applied almost exclusively in cases of sexual abuse (but could be applied in other settings as well: Greuel, 2001: 6; Niehaus, 2008: 318–19; Arntzen, 2011: 1–2) is the above mentioned *Statement Validity Assessment.* In the 1960s the psychologist Udo Undeutsch integrates the studies realized up to this time into his model of the so-called "Realkennzeichen", reality criteria or truth criteria (Undeutsch, 1967: 127–56). This approach is based on the obvious, common sense-based assumption (cf. Undeutsch, 1967: 126), that verbal structures of true and false statements differ significantly ("Undeutsch-Hypothese", cf. Undeutsch, 1967: 125–26). Note that in fact this hypothesis is basic and is shared, more or less implicitly, by the supporters of the Reality Monitoring and SCAN techniques (cf. Vrij, 2009: 261–289), as well as explicitly by the authors of the LIWC[6] software, a lexicometric approach to lie detection: "Although liars have some control over the content of their stories, their underlying state of mind may 'leak out' through the way that they tell them" (Newman et al., 2003: 665).

Not surprisingly, these three techniques define criteria that partly overlap with Undeutsch's truth criteria like *affective information* (LIWC; the core criterion in RM) or *spontaneous corrections* (SCAN), with the latter aspect however interpreted as cue for lying in SCAN. For assessing the truth of statements, Undeutsch suggests analyzing its "content" taking into account linguistic and rhetoric features and the content structure as a whole. As a consequence, the psychological expert's attention is explicitly drawn to the testimony, away from the witness' general trustworthiness.[7]

Undeutsch's model fits well in the German inquisitorial (and not adversial) trial system that focuses on establishing facts and relies on expert judgements and less on arguing in the court room. Therefore the *Realkennzeichen*, and their revised version by Steller & Köhnken (1989: 221)[8] have been and continue to be widely applied.

---

[6] LIWC is the abbreviation for *Linguistic Inquiry and Word Count*, a tool used for scientific purposes; also ch. 4.

[7] Note that the trustworthiness of children, especially of girls, has long been denied (Undeutsch, 1967: 26–27).

[8] There are alternative revisions of Undeutsch (cf. Arntzen, 2011: 17–19), like the commented and enlarged criteria list proposed by Arntzen (2011: 25–98).

Table 1: Content Criteria.

GENERAL CHARACTERISTICS	1. Logical Structure
	2. Unstructured Production
	3. Quantity of Details
SPECIFIC CONTENTS	4. Contextual Embedding
	5. Descriptions of Interactions
	6. Reproduction of Conversation
	7. Unexpected Complications during the Incident
PECULIARITIES OF THE CONTENT	8. Unusual Details
	9. Superfluous Details
	10. Accurately Reported Details Misunderstood
	11. Related External Associations
	12. Accounts of Subjective Mental State
	13. Attribution of Perpetrator's Mental State
MOTIVATION-RELATED CONTENTS	14. Spontaneous Corrections
	15. Admitting Lack of Memory
	16. Raising Doubts about One's Own Testimony
	17. Self-Deprecation
	18. Pardoning the Perpetrator
OFFENSE-SPECIFIC ELEMENTS	19. Details Characteristic of the Offence

Note: Adapted from Steller & Köhnken (1989: 137).

3.2. The SVA as a Combined Technique

The content analysis with focus on verbal elements is the central part of the statement validity assessment (Greuel, 2001: 261), but SVA equally gives weight to non-verbal factors like age, narrative capacity etc. This integrated assessment technique is accepted as good standard, especially in cases of abuse, through a ruling of the German Supreme Court in 1999;[9] it is also accepted at court in some North American contexts (Vrij, 2009: 201) and in various other European countries (Volbert & Steller, 2014b: 207).

In spite of its success the Undeutsch based statement analysis has undergone terminological and methodological adjustments. As reported by one of the most authoritative experts, Luise Greuel (2001: 26), the prescientific and somewhat moralistic expression, "credible", has been substituted – at least in psychological context – by the more adequate term "erlebnisbasiert", 'experience-based/related', with the sometimes used opposite term "fabricated" (Volbert & Steller, 2014b). As far as the method is concerned, Greuel (2001: 16–20; see also Steller & Köhnken, 1989: 217, Daber, 2014: 263–64) identifies three main evaluation attributes of psychological SVA:

---

[9] See the decision of the *Bundesgerichtshof*, BGH 30.7.1999, 1 StR 618/98.

1. Eye-witness ability ("Aussagetüchtigkeit"), i.e. the general ability to report personal experiences; the general narrative capacities. This is of high relevance in cases of sexual abuse, as Volbert & Steller (2014a: 416) point out: only the children's ability to present consistent narratives on their own will rule out the risk of manipulation through the interviewer's encouraging questions.[10] Though narrative capacity is no content-related aspect, Undeutsch proposes its inclusion for veracity assessment (1967: 56; 1984: 51). In addition, as far as children are concerned, there has to be kept in mind the "lying competence" as well. Lying, as defined above, requires being able to identify intentionality of human behaviour, a skill acquired at the age of about 10 years (Greuel, 2001: 233); some findings however suggest the ability to deceit 'efficiently and firmly, even when asked with some pressure' at the age of 7–8 years (Volbert & Steller, 2014a: 416: "effektiv und standhaft, also auch bei Nachfragen").

2. Quality of testimony ("Aussagequalität"), i.e. the content and structure of the testimony, usually assessed with the help of Steller & Köhnken's criteria list to evaluate the experience relatedness (Content Based Criteria Analysis, CBCA), taking into account, especially when assessing testimonies of children, the individual baseline. The quality could as well be assessed with respect to the intraindividual consistency of content in various statements regarding the same crime (Arntzen, 2011: 52–53; Volbert & Steller, 2014c: 396).

3. Reliability of testimony ("Aussagezuverlässigkeit"), i.e. the absence of all factors that might lower the witness' reliability, like mental diseases or externally induced "false memories" (Greuel, 2001: 19–20). In addition to the sources of non-experience-based reports as discussed in section 2, it is especially the phenomenon of "false memory" that has been object of some recent research work, as Volbert (2014b: 208) and Daber (2014: 259) point out. Memory of non-factual abuse might be insinuated in cases of divorce by the witness' mother. Note that the relationship between child and mother is one of absolute loyalty – even more so, if the mother is the only trustworthy person left to rely on. The child quite will probably submit the non-factual experience due to what is called "Quellenvermischung" ('confusion of sources', Volbert, 2008: 336), but he or she will not behave linguistically as a liar since she (or he) is not aware of or does not intend to deceive (Welle et al., 2016: 118). The language of false memories ("Pseudoerinnerungen", Volbert & Steller, 2014c: 391) therefore only slightly differs from spontaneous experience-based narrations, containing for instance slightly less details (Volbert, 2008: 339). In fact, as one study with preschool-children shows, experience-based and false suggested statements indeed become "increasingly similar over the course of repeated interviews" (Volbert & Steller, 2014b: 215).

There is, in addition, an overall trend towards dissolving the clearcut distinctions between fiction and reality due to intense media consumption (cf. Daber, 2014: 260). These

---

[10] The authors of the RM approach point to some age-related issues as well, they cite findings concerning children's missing ability to distinguish memory of own words from memory of own thoughts (1981: 76).

sources of unreliability should, following Undeutsch's axiom about the distinguishability of truth and untruth in language, make for the absence of linguistic differences. In practice, the hypothesis of any type of induced experiences is evaluated before the analysis of language.

The first and the third step of analysis correspond – roughly – to what Vrij (2005: 7) calls analysis of "external factors", that is of non-linguistic features, to be taken into account by the expert for the final interpretation of the statement.

Linguistically relevant is the second step of the witness' exploration, i.e. the analysis of the quality of the statement ("Aussagequalität"), that includes the application of the by now classical content-criteria. The German psychological experts currently refer to the revised Undeutsch version, see above, table 1. While Undeutsch presented 15 criteria, Steller & Köhnken list 19 criteria, due to more fine-grained distinctions.[11] Undeutsch for example quotes the describing of the 'development of the relationship' (criterion 9, "Entwicklung der Beziehung", cf. Undeutsch, 1967: 149–52) as a valid criterion for veracity. Steller & Köhnken propose relying on more concrete cues instead: "descriptions of interactions", "reproduction of conversations" (which of course are just a partial substitution of Undeutsch's criterion 9). By the way, none of these two cues is part of the RM or SCAN list; the second cue however is considered in the study of Adams & Jarvis (2006) as truth-related.

3.3. Testing Reliability of SVA and CBCA

Since their publication the criteria do work very well in practice. They seem to be, if applied by trained experts, of relatively high reliability and have good predictable accuracy rates. The average accuracy rate is already 70% in laboratory and field studies, as cited in Vrijs manual (2009: 235).[12] Since complete SVA evaluations of witness statements are performed by psychological experts and are based on more criteria than only those referring to content, there might be an accuracy rate above 70%, as the Canadian psychologist John Yuille states already in 1989, referring to the methods of *Aussagepsychologie*: "Statement analysis techniques however, have proven both reliable and valid in assessing the credibility of children's statements." (Yuille, 1989: X). Wegener (1989: 127) underlines that "experts usually agree in their judgements" when applying the content-criteria. In 2005, Vrij presents a review of the first 37 laboratory studies which test the Steller & Köhnken criteria. Again, the "interrater agreement" turns out to be "adequate" as far as the criteria are concerned, but less convincing when referring to the statement analysis as a whole (Vrij, 2005: 34). Interestingly, the authors of a more recent meta-analysis (Hauch et al., 2017) recommend using the criteria-analysis *only within* the SVA

---

[11] For a concise review of different re-elaborations of the *Realkennzeichen* see Steller & Köhnken, 1989: 218–221.

[12] In 2005, Vrij reports an average error rate of 30% in laboratory studies (Vrij 2005: 32).

(Hauch et al., 2017: 828; in accordance with Daber, 2014: 261), after, however, having confirmed Vrijs finding, i. e. good interrater reliabilities with respect to 15 CBCA criteria especially in field studies, less in laboratory experiments (Hauch et al., 2017: 825). According to Vrij (2005: 33) the results of pertinent studies suggest that statement analysis does not yet meet the US-American conditions determining a method's acceptability for use in court, as partly confirmed by Hauch et al. (2017: 826).

3.4. Testing Validity of SVA

Steller (1989: 142) proposes two main types of possible testing of any criteria validity: Prospective and retrospective field studies and tests with simulated lying. Especially the latter type is extremely popular in veracity research, typical examples could be DeCicco & Schafer (2015, testing SCAN criteria), Almela, Valencia-García & Cantos (2013, testing LIWC criteria), while field studies are still rare, as regretted by Welle et al. (2016: 112) and Volbert & Steller (2014b: 210); reservations have to be put forward for both techniques. The main and up to now unsolvable problem for field studies is the missing control value, that is there is no way to identify objectively the truthfulness (149), since the ground truth (Fitzpatrick et al., 2015: 2; Vrij, 2005: 8–9) can't be established. On the other hand, cue-tests with simulated lying yield heterogeneous results, probably due to the varying topics (Almela et al., 2013: 9). In sum, simulated lying concerning the topic *abuse*, then, seems impossible for research purposes, due to ethical reasons (Steller, 1989: 145; Vrij, 2009: 220).

Undeutsch had his reality criteria tested, with convincing results, as he himself reports (1967: 171) – without providing any data though; in 1984 he claims for the up to then about 50.000 SVA reports[13] that there is "not a single case to be found [...], which later, in the criminal proceeding or afterwards, turned out to be in conflict with other relevant evidence." (1984: 64). Hauch (2017: 829) however points out, that there have well been cases of expert conflicts. Volbert & Steller (2014b: 210), reviewing recent findings, can quote positive results for reliability, but they emphasize that the identifiable differences between experience-related and fabricated statements most probably will be more distinctive in real forensic than in artificial laboratory settings. This hypothesis is confirmed with impressive statistical values by a psychometrical analysis (Steck et al., 2010, s.p.), performed with content criteria. The authors contrast true and not-true laboratory-created with true and not-true authentic statements, demonstrating the significantly higher discriminative value of content criteria for authentic statements; similar findings are quoted by Fitzpatrick et al. (2015: 41), concerning one of the most important SCAN criterion ("statement balance").

---

[13] In 2017 the number of cases amount to about 100.000 (Hauch et al., 2017: 820).

In spite of promising results from field studies, SVA experts as a rule are reluctant to subject their method to *testing*, without, however, excluding research in general. With respect to statement validity assessment, they emphasize that what decides the outcome of a complex, intraindividual and case-specific interplay of different factors is the expert's final judgement on experience-relatedness. So, if these factors, taken separately, have no attributed constant value and do not constitute a "checklist", their validity cannot really be tested or at least tests have to be considered with a good deal of scepticism (Arntzen, 2011: 11; Steller, 2009: 303; Niehaus, 2008: 315; Undeutsch, 1984: 63–64).

3.5. SVA Criticism

Despite the undisputed success of Undeutsch's and Steller & Köhnken's content criterialist in practice, as part of SVA (cf. Undeutsch, 1984: 64), there is widespread dissatisfaction.

Already at first sight one may notice an inconsistency: not all content criteria explicitly refer to verbalized *content*. While criteria 3–10 and 19 do in fact refer to the information given in the testimony, 1, 2, and, in a way, also 14 account for structural aspects of the testimony, 12 as well as 15 to 17 concern a change in the narrator's perspective. Regarding their degree of identifiability the criteria differ as well: compare criterion 4 "contextual embedding" to 6 "reproduction of conversation". While the latter is linked to predefined formal verbal cues (e.g. *verba dicendi*), the former criterion refers to individual, case specific features. This inconsistency in the criteria list has also been noticed by Steck et al.:

> "The term 'experience-basedness' is supposed to be the category on which the reality criteria are based. However, the concrete definition of the reality criteria refers to very different aspects of the narrative and reporting style […], with the result that any sense of homogeneity of the reality criteria must remain problematic despite a semblance of unity of the seemingly unequivocalness of the construct of 'experience-basedness'."[14]

Unsurprisingly, the different criteria "vary widely with respect to the precision with which they are operationalized" (Hauch et al., 2017: 820). But, again, recall that as part of the integral statement validity assessment and when founded on transcripts of considerable length, they work successfully.

The heterogeneity of the content criteria is a consequence of the absence of a consistent theoretical underpinning, which is in fact what has been criticised explicitly and repeatedly. In 1992, Steller, Wellershaus & Wolf summarized:

> "There is still no answer to the questions how the content features of true statements are founded theoretically and how they could be linked to psychological theories and concepts."[15]

---

[14] Steck et al., 2010: 4 – translated from German: "Für die Realkennzeichen wird mit dem Begriff der Erlebnisbegründetheit ein homogen erscheinendes Konstrukt reklamiert, bei der Ausformulierung der Kennzeichen wird aber auf unterschiedliche Funktionen des Berichtsstiles Bezug genommen (Steller & Köhnken, 1989), so dass die Homogenität der Realkennzeichen trotz des eindeutig anmutenden Konstruktes der Erlebnisbegründetheit offen erscheint."

34

In 2014 still "not all criteria are theoretically well backed" (Volbert & Steller, 2014b: 212) and in 2016 Welle et al. still criticize the lack of any scientific procedure to determine ground truth (2016: 112).

Greuel (2001: 347) goes further and argues for an inclusive turn (*integrative Wende*) in statement analysis. She claims for a comprehensive theory of statement in general, proposing to integrate forensic and sociological parameters; a contribution by linguistics is, unfortunately, not considered.

As far as empirical work is concerned in 1989 Steller wonders why, in spite of an "enormous number of cases" there are "no scientific reports" (142). The studies that have been conducted in the last two decades seem to be of no help: there is a "large heterogeneity between individual studies for each CBCA criterion" regarding interrater agreement (Hauch et al., 2017: 824), in part due to topic- and genre-variation.

Another, more specific reason of discontent is the lack of any scientifically based weighting of the criteria (Welle et al., 2016: 116–117; Steller, 1989: 39; Wegener, 1989: 128), while in practice the psychological experts of course might have subjective preferences when assessing credibility. There is, however, some indirect weighting, i.e. the role of necessary conditions has been attributed to the first three criteria (cf. Hauch et al., 2017: 820; Undeutsch, 1984: 60), at least to the first two (Steller, 1989: 136).

A further point of criticism is the unidirectionality of the SVA approach (Niehaus, 2008: 315) that provides criteria for truth-identification only. This may be due to a general human truth-bias (cf. Welle et al., 2016: 118; Steck et al., 2010: 14) and due to the fact that the experts' reports are to be presented at court, towards judges and lawyers who essentially do not want to know what is *not* true, but what *is* true, i.e., on what they can base their judicial narrative as the facts on which the legal norm to decide the case is to be based. Hettler (2005), however, has suggested enlarging the CBCA by adding empirically confirmed *deception* cues (Hettler, 2006).

So, in sum, it can be said that (cf. Vrij, 2015) the application of the instruments discussed here in their present state needs to be realized with caution, depending in particular on the evaluation of external evidence. This applies the more as long as verbal parameters are not further elaborated and applied with more technical precision, and as long as there are so few verbal parameters. This applies in the same measure, if not more, to automated tools.

---

[15] Steller, Wellershaus & Wolf 1992: 167 – translated from German: "Unbeantwortet ist jedoch nach wie vor die Frage, welche theoretischen Annahmen den aus gerichtlicher Praxis abgeleiteten inhaltlichen Merkmalen wahrer Bekundungen zugrunde liegen und mit welchen psychologischen Theorien und Konzepten sich diese Ansätze verknüpfen lassen."

4. Linguistically Oriented Research on Lying

Empirically oriented studies are realized in computational lexicometric analysis, often based on the LIWC model (see also section 3.1; Pennebaker et al., 2001) that includes 29 potential cues for deceptive language. The results of five separate studies suggest that five out of these 29 seem to be good indicators for lies, with lie detection rates of about 70 % in laboratory studies: the low frequency of first-person pronouns and of restricting conjunctions ("exclusive words") like *but*, and high frequency of third-person pronouns, motion verbs and negative emotion verbs (Newman et al., 2003). The latter cue has recently been confirmed by Almela et al. (2013) in a study with Spanish speakers. Taylor et al. (2017) employ LIWC to explore culture-relatedness of lying behaviour.

More theoretically oriented discussions usually locate lying and deceiving as an issue in (cognitive) pragmatics. According to Meibauer (2014: 103; see also: Dietz, 2003: 25; Galasiński, 2000: 23) lying means consciously asserting the untruth with the intention to make the hearer assume the assertion to be true. Lying by using language is considered a subclass and a prototypical form of deceiving as reported in Galasiński (2000: 18), Meibauer (2014: 25) and especially in Horn (2017: 31) for a linguistically watertight distinction between deception and lying.[16] The fact that lying is considered as the morally most reprehensible form (Dietz, 2003: 7) is an important aspect, insofar as it underlies the "cognitive load" hypothesis for lying, as this additional burden will be reflected in additional formulation work, reflected in deviating types of language. Lying always implies breaking the Gricean maxims (Meibauer, 2014: 76–79; Galasiński, 2000: 98) and seems to be linked to some constant cues: the meta-study by De Paulo et al. (2003) cites 158 different cues mentioned in the reviewed articles, about two thirds of them are language-related.

Smoother strategies of deception, discussed and subsumed under *evasion strategies* in Galasiński (2000) are connected to dialogical, highly reciprocal situations with two equal partners, where turntaking develops more or less freely. In fact, Galasinki concludes that "evasiveness is an attribute of an utterance used as a response" (111).

Lying in between the lines, e.g. by falsely implicating (Meibauer, 2005: 1384; 2014: 140; Horn, 2017) equally comprises just a single turn in conversation, conversationally just a small portion within a larger dialogical context. In veracity assessments the conversation is most likely unidirectional, with one speaker interviewing and the second speaker answering and providing, hopefully, informative, almost monological, narrative contributions. Deceptive tactics like evasive utterances and false implicatures are unlikely in such situations of psychological assessment, as the least interpretative ambiguity would be noticed and would not be accepted by the exploring psychologist. Protoypical lying is, by contrast, a demanding cognitive task and is hypothesized to show at the linguistic

---

[16] In contrast to these approaches, Vrij (2009: 15) considers both verbal *and* nonverbal cues, using *lie* as general term.

surface, at least in cultures where lying is morally burdensome, as in Western cultures. Psycho-neurological studies empirically measures behavioral neurological parameters and interpret them with respect to processes hypothesized to go on in the person's mind. Suchotzki (2015: 395) measure processes in particular brain potentials presumed to reflect processes of anticipation and response preparation. She can provide evidence that lying indeed needs more preparing time due to considerably higher cognitive efforts, supposed to result in noticeable, deviating linguistic behavior. It is a moot point to which extent such treacherous additional cognitive work, that gives it all away, is triggered only by the prototypical lie, uttering propositions that are knowingly not true with the intent to deceive, or also by the many in-between cases such as discussed in Horn, 2017 (from withholding information to the "reservatio mentalis"). Obviously, previous empirical research has focused on the prototypical lie.

At this point another crucial issue concerning baselines comes into play. What was termed "noticeable linguistic behavior" or "deviation" from some expected linguistic norm implies the notion of an identifiable "baseline" or a notion of "normalcy" in language use. And "identifiable" in turn implies that it can be identified by established linguistic methods, and not just by intuition. This normalcy assumption can be of two related kinds: the occurrence of a linguistic structure (like "self-correction" or "motion verb" – both to be discussed below), or the quantitative distribution of some linguistic form in a corpus. The key issue in linguistic research is the baseline against which the "deviating" form registers, or is manifested in a statistically significant deviation. This issue takes several shapes and is discussed as a methodological issue in other places. What must be emphasized in the present context is the relativity of the deviating structure. From the linguistic point of view it is a paramount methodological flaw to assume, as already pointed out initially (section 1), that given structures are invariably associated with a cognitive process, such as lying.

Some interesting insights might be gained when narrative, transclausal structures (as determined by Labov & Waletzky, 1997 or Quasthoff, 1980; see also Greuel, 2001: 330–332) of a longer extension are included in the statement analysis. Structuring devices like discourse markers ("well", "now") or tense changes for instance, well documented for experience-based narratives, could show different distributions in false statements, or not occur at all. Tense-changes for instance are clear cases of evaluative, emotional comment common in narratives, and not always and in themselves suspicious or indicative of extra cognitive work in order to produce a full-fledged lie. Instead they are a normal part of emotional evaluation and peak marking. In their study on written statements presented to the police, Adams & Jarvis (2006: 18) demonstrate that liars typically produce longer prologues in their narratives than truth-tellers. It has to be kept in mind however, that also the truthful report of experienced traumatic events like sexual abuse might not obey baselines, i.e. common narrative organisation priniciples (Greuel, 2001: 85); the recall of such events, however, seems to work well and in the ordinary manner (Vrij, 2009: 221).

The argumentative interpretation of linguistic forms in a rendering of past experience therefore depends on where a true narrative is offered, and at what point in the narrative the form occurs. The presence or absence of forms must be interpreted relative to the structural part of the narrative. If the narrative report is offered by adolescents or children, it is an issue to what extent early age groups already have a "narrative" competence such that a "normalcy" expectation for the occurrence of the forms in question makes sense.

The complexities in interpreting tense changes – the individual occurrence or a statistical pattern – become clear when tense changes are not interpreted with respect to their normal function and occurrence in narratives. For instance Olsson (2004: 132f.) interprets a change to anterior tense ("had not said") as to "set up the narrator for his later denial" about the use of a gun – when it is simply a case of providing static background information in the course of a complicating action series with simple past tenses. The tense changes in text 9.8 (Olsson, 2004: 133) are clear cases of evaluative, emotional comment common in narratives, and not in itself suspicious or in themselves indicative of extra cognitive work in order to produce a lie. Instead they are a normal part of emotional evaluation and peak marking,

Just as, from a developmental point of view, it is an open question whether children can "properly" lie in terms of empathy development, the issue is also from what developmental age children and adolescents (and on occasions also adults) can master syntactic structures that are unusual or "marked", such as extrapositions, inversions, preposings or dislocations. The absence or presence, or quantitative deviation, of these structures that are candidates to be considered as indications of cognitive load in lying must always be interpreted in the light of baselines that take in account genre, age, linguistic nativeness and related factors.

5. Verbal Cues of Lying: Challenges and Desiderata

In the preceding section we pointed out that one of the key issues in establishing baselines, and in creating corpora that allow forensic argumentation on a qualitatively fine-grained quantitative basis are the notion of the genre, and constrained by genre, topic. It is obvious that the presence, absence or relative length of one section in a statement is crucially dependent on the larger genre in which these sections are embedded: in a true narrative with larger monologic parts or in an interactive question-answer dialogue.

The crucial role of the genre for the occurrence of which type of cue can be further illustrated by a positive example of the application of linguistic cues in a deception detection program just implemented, *VeriPol,* a program to detect insurance fraud by people reporting e.g. expensive cellphones as stolen or being robbed of them (Quijano-

Sánchez et al., 2018). The discussion up to now implied a relatively "open" and "unconstrained" world of discourse, where in principle there are no or very few restrictions on what kind of situation and topic occurs and, consequently, what kinds of objects and events can be expected in the text. The program VeriPol uses as linguistic cues from a situationally highly constrained situation, with the strong likelihood of occurrence of a limited range of lexical items (e.g. "seguro", "iPhone", "contrato", "mochila", "bolso", "casco", "negro", "no puede dar más datos", "no puede reconocer") – items that are highly likely and predictable to occur in such a discourse where a robbery is falsely reported to the police in order to claim compensation by an insurance. This type of linguistic cue, together with others from grammar, is used with a 91% success. Again, even a tool with such an unusually high success rate is used only not as *the*, but as *one* indication of a criminal act. It is a "decision support system", and not a categorical decision maker.

If genre-baseline in combination with possible individual verbalizing preferences is one of the key linguistic challenges in forensic work, the other is the definition of the linguistic categories that are supposed to be diagnostic. The current state of affairs is that the results of empirical testing of linguistic diagnostic cues are in themselves inconclusive and often contradictory (Hauch et al., 2017: 826), and are therefore not yet a good method to lead to improvements of the psychological techniques discussed in sections 3 and 4.

Categories used such as "equivocation" (Adams & Jarvis, 2006) lump together types of expressions that linguistic analysis would clearly treat as very separate things, such as "maybe","roughly", or what linguists call "hedges". Other cases in point are categories like expressions of emotion that often enough subsume subtypes that express very different types of content. For instance, the appearance of direct speech in the peak section of a narrative is an element of emotional evaluation, but would not be classified as the expression of emotion by most studies. While LIWC authors Newman et al. (2003: 672) categorize the use of "negative emotion verbs" as symptom of lying,[17] in SVA all reporting "subjective mental state" (crit. 12) is explicitly interpreted as symptom of experience-relatedness – two contradictory interpretations of the same phenomenon. Greuel (2001: 100) however, quoting findings from cognitive psychology, adds that reporting of true, experienced, negative feelings probably is less detailed: this indirectly supports Newman's view.

Some other categories are not yet exploited at all. Adverbs, for instance, apart from being notoriously polyfunctional (like "now", whose temporal meaning is only one of many others) often do many things at the same time, like functioning as a discourse marker announcing another segment of discourse. Another notorious, and completely underexploited (at least in forensic linguistics) class of expressions, are particles ("well", "so"), which can provide access to deeply embedded beliefs and presuppositions, but are hard to handle from the point of view of linguistic analysis. They nevertheless could serve as linguistic access to content that is normally well below the threshold of awareness and manipulability, and therefore of specific value.

---

[17] This remains unchanged in the updated version LIWC2007, see liwc.net/LIWC2007LanguageManual.pdf.

So the other key challenge from the side of linguistics is to refine and further differentiate the linguistic categories, in order to be able to build degree zero corpora that can be fruitfully used as baselines.

In judicial practice experts can't do without cue lists. But these cues need more serious scientific underpinning from functional theories of language use as well as from empirical research based on reliably established ground truth. This will lead not only to improved validity scores but also to clearcut, therefore easily applicable definitions.

6. Examples

This chapter discusses two diagnostic linguistic cues as far as they are identifiable in extracts of four authentic interviews, belonging to completed, sentenced abuse trials. The four transcribed explorations have been conducted by a psychological expert with female victim-witnesses. One testimony has been assessed as not experience-based while the other three all have been independently assessed as true. All quotations in the following are taken from anonymized versions of the extracts. For the purposes of the present publication, some lexical elements that are not part of credibility cues, have been slightly altered. The extracts mainly contain narratives concerning the alleged offence, but they are in fact mixed with other types of discourse, one of the characteristics of the data in field studies (Fitzpatrick et al., 2015: 58), so some interindividual observations (see ex. 8, 9) are possible.

If we consider research done in the field of deception detection, we find some similar linguistic cues that are discussed with respect to their truth discriminative power in quite different research contexts. Two of these core cues, *motion verbs* and *self-corrections*, will be discussed more closely in the following.[18]

6.1. Motion Verbs

The widely used tool *Linguistic Inquiry Word Count,* (LIWC, Pennebaker et al., 2001) enables the analysis of the distribution of 29 lexically manifest cue-categories in written English texts, e.g. "past tense verbs", "causation". In 2003, Newman et al. published the results of their own LIWC laboratory applications: the category "motion verbs", comprising the selected verbs *to walk, to carry, to go,* belongs to the five 5 LIWC categories that seem to have significant predictive qualities, being typically used in the false statements (Newman et al., 2003: 672). This seems to be inconsistent with findings in psychology of

---

[18] Fobbe (2018) has presented a similar linguistic analysis of the LIWC criterion "use of first/third person pronouns", where third person pronouns are supposed to be more frequent in lies.

perception, as reported in Greuel's manual: Motion events are perceived with more ease as real and are, so far an implicit conclusion, recalled much easier than static and abstract contents (Greuel, 2001: 50, notes: "Eigenbewegt erscheinende Konfigurationen werden eher als wirklich angesehen als passiv bewegte oder ruhende Reizkonfigurationen"). The description of motion therefore is to be expected in experience-based statements. Finally, since motion verbs explicitly locate events in the situational context, especially if combined with indicators of direction, like e. g. *to rush out*, they must be considered as one concrete realization of the quite vague SVA and Reality-Monitoring criterion "contextual embedding" / "spatial information" (Vrij, 2009: 267) – as a credibility cue, again. Motion verbs are no criterion in SCAN.

Motion verbs could be retrieved automatically without greater problems. Their distribution in the whole text, their grade of semantic specificity, their – up to now never analysed – relation between types and tokens could be graphically represented.

But things are not as simple as that, as the authentic examples show. See for instance:

(1) (established false statement)

Question: "Ist dann noch irgendwas passiert oder dann nichts mehr?"

'Did anything happen after that or nothing else?'

Answer: "Nein, dann nichts mehr. Dann hat er die Hose wieder angezogen, also zugeknöpft, quasi, hochgezogen, zugeknöpft und dann bin ich mit ihm eh rausgestürmt"

'No, nothing else. Then he [not the presumed perpetrator, but another presumed, male, victim; MN] put on his trousers again, that is he closed the buttons, more or less, pulled them up, closed them and then we rushed out'

What is surprising here is the semantic specificity of the motion verb, in the same line as *walk*: "gestürmt"/ 'rushed' ("raus-"/ 'out' indicates the direction, called *path*). There are no such exact indications of manner of motion in the three experience-based extracts, but there is still a lot of motion-coding, realized mainly by the much less specific motion verb *gehen*, meaning in these contexts 'to move', plus *path*-indicating satellites like *weg*, *hin*, *rein* ('away', 'there', 'into') etc. especially in the description of sexual activities. This frequent type of motion expression in true narrations seems to be a veracity cue of considerable weight.

Obviously the research work on the use of motion verbs, and in general on contextual embedding, in testimonies has to be refined. The present data suggest that the descriptions of motion in true statements point at coding *path* with the help of adverbial satellites and neglect *manner*. That is, the verb *to go* in the text data of Newman et al. (2003) might behave quite differently with respect to truth as the more specific manner-verb *to walk*. Probably it is the heterogeneity of Newman's motion verb category that has caused its poor performance (within the five best performing criteria) in Newman's own test, but also e. g. in a study with written Spanish data (Almela et al., 2013: 8).

The blanket criterion "motion verb" is misleading – still arguing on the basis of the present data. It is the different codings of the major semantic category *motion* that is

interesting instead. The combined coding of the subcategories *manner of motion* and *path* (*to rush out*) has to be distinguished from the coding of just *motion* and *path* (e.g. *to go in*). In informal spontaneous narratives in German the latter combination, the so-to-say *path*-emphasizing combination is preferred and can even be verbalized without motion verbs, just by modal verbs or the copula *sein* ('to be') plus path-indicating satellites, i.e. directional adverbs or prepositions, like in the following examples taken from the extracts of experience-related statements: "in [...] rein wollte" ('wanted into [...]') or "is die ganze Zeit über mich" ('is over$_{directive}$ me the whole time'). These two examples seem to be elliptical, i.e. seem to imply the verb *gehen*. Both are without doubt coding *motion* and *path*.

Example (1), by the way, shows another structure that is a candidate for linguistic cue for lying: "quasi". The best translation would arguably be some postposed *more or less* or *like*. It would be correctly classed as an indicator of qualifying one's statement one does not wish to commit to, insofar leading over to the next class of expressions.

6.2. Corrections

This criterion belongs to the CBCA lists as criterion 14 in Undeutsch (1967) and Steller & Köhnken (1989) and is considered uniquely as typical of experience-based statements. Greuel (2001: 35) argues that corrections don't serve the aims of liars, who prefer to create an impression of absolute reliability: "Täuschende wollen erinnerungssicher wirken."

It is not part of Reality Monitoring, but it figures in SCAN, referring to crossed out segments in written data only. Naturally corrections can't be retrieved by a lexicometric tool like LIWC, therefore they are not part of the LIWC list of cues.

Undeutsch (1967: 162) subsumes under this point amendments of previous statements ("Verbesserungen") as well as specifications within an utterance ("Präzisierungen") and, quoting another study, immediate self-corrections ("plötzliches Sichselbstverbessern"). Vrij (2009: 212) translates as "correction" (within a sentence) or "addition", or even "explanations" (Vrij, 2009: 284). Niehaus (2008: 313), in her rearrangement of the CBCA criteria, prefers to speak of spontaneous specifications and corrections ("spontane Präzisierungen und Korrekturen") to be subsumed under her so-called motivation-related aspects ("motivationsbezogene Aspekte").

All these terms – amendment, specification, correction, crossing out – are treated as synonyms in the above quoted publications and refer to different types of what is called in conversation analysis "self-initiated self-repairs", that is they refer to not induced ("immediate") utterances of the interviewee, always put forward to improve previous utterances and parts of utterances in some way. Not surprisingly, Arntzen (2011: 19) ranges this vague criterion among the problematic CBCA criteria of just 'impressionistic value'

("Eindrucksqualitäten") and adds, that "Verbesserungen", 'amendments', are as common in true as in untrue statements. In Vrijs and Hauch's meta-studies the criterion "spontaneous corrections" is listed under the criteria with unsatisfying or insignificant interrater reliabilities (Vrij, 2005: 16; Hauch, 2017: 826). The following examples, all good candidates for this criterion, show confusing formal differences that might explain its weak reliability:

(2) He wore black trousers, no, sorry, they were green.

(3) We were in his car [...] by the way it was a Volvo [...]

(both from Vrij, 2005: 212; fictional illustrations for criterion 14)

(4) (offense-related, established true statement)
Da hat er meine Hand so gekriegt – gerissen richtig,
'Then he caught my hand – really tugged at it,'

(5) (offense-related, established true statement)
Die Aussagen halte ich [...] aufrecht. Ich habe lediglich zu berichten, daß [...]
'I confirm [...] my statements. I just have to report that [...]'

(both from Undeutsch, 1967: 152; authentic material)

Some of our own authentic material (extracts) contains examples such as:

(6) (offense-related, established false)
[...] hab ich zum NAME1 gesagt, also zum großen NAME1, zum SURNAME [...]
'[...] I said to NAME1, I mean to the older NAME1, to SURNAME [...]'

(7) (offense-related, established false statement)
[...] daß er den unter Druck gesetzt hat, eingeschüchtert hat.
'[...] that he put pressure on him, intimidated him.'

 (8) (not offense-related, individual "baseline", established true statement)
Also ich war ich bin halt so mit, also ich war bei meinem Onkel.
'That is, I was, I just went with him, that is I was at my uncle's place.'

(9) (offense-related, established true statement)
[...], dass ich also erst hat es angefangen mit den Händen.
'[...] that I, I mean, first it began with the hands.'

Four main types of "spontaneous corrections" could be distinguished, three of them are repairs:

a. Content substitution, i.e. verbalized "crossing out" (2);

b. content specification, realized as inserts and introduced by the editing terms *by the way, also* (3, 6) or realized as addition of partial synonyms, not introduced (4, 7);

c. changing of syntactic planning, introduced by the editing term *also* (8) and not introduced (9);

d. (no repair) content addition, realized in a new statement and announced (5).

So the notion of "correction", tempting as it is to be applied in a blanket way in forensic analysis, needs much more differentiation in terms of what these structure do and how they are caused in discourse. It is therefore too early to draw any conclusions concerning

the veracity-discriminating power of each of the four main correction-types. It has to be kept in mind, however, that repair strategies differ as far as their cognitive load is concerned: syntactic repairs seem to be typical for relaxed talk, with little control (8, 9). Content repairs however concern lexical elements, follow therefore a specific idea of what has to be represented and how content needs to be semantically and lexically packaged in order to sound convincing – arguably with more refining work required and creating more tension in the case of lying.

7. Conclusion

We have argued, in line with criticism, that the current procedures in evaluating veracity in verbal report by witnesses should be not supplanted, but complemented by linguistic technical knowledge of a specific kind. This appears called for several reasons. A better, professionally linguistically informed understanding of the functioning of communication through contextually and functionally fully embedded communication through language is essential for formulating hypotheses about which linguistic structures can be reasonable, more or less "weighty", candidates for diagnostic status in veracity evaluation, i.e. reliable evaluation of experience-based as well as fabricated statements. The formulation of such hypotheses can be based on a much more fine-grained analysis of the nature of specific linguistic expressions, always bearing in mind their functions and privileges of occurrence in differentiated genres and their cultural embedding. This addition of applied linguistics to the toolbox will lead to a much more successful procedure and a much more differentiated notion of baselines – if carried out in conjunction with other, psychology-based approaches.

References

Adams, Susan H. & Jarvis, John P. (2006). Indicators of veracity and deception in analysis of written statements made to police. Speech language and the law. *International Journal of Speech, Language & the Law*, vol. 13, 1–22. DOI: 10.1558/sll.2006.13.1.1.

Almela, Angela, Valencia-García, Rafael & Cantos, Pascual (2013). Seeing through deception: A Computational Approach to Deceit Detection in Spanish Written Communication. *Lesli*, vol. 1, 3–12. DOI: 10.5195/lesli.2013.5.

Arntzen, Friedrich & Michaelis-Arntzen, Else (2011). *Psychologie der Zeugenaussage. System der Glaubwürdigkeitsmerkmale*. Beck.

Becker, Tabea (2005). The role of narrative interaction in narrative development. In Uta M. Quasthoff & Tabea Becker (Eds.). *Narrative interaction* (pp. 93–111). Amsterdam: Benjamins. DOI: 10.1075/sin.5.05bec.

Daber, Beate (2014). Neue Entwicklungen in der Aussagepsychologie. In Rüdiger Deckers & Günter Köhnken (Eds.). *Die Erhebung und Bewertung von Zeugenaussagen im Strafprozess. Juristische, aussagepsychologische und psychiatrische Aspekte* (pp. 259–66). Berlin: BWB Verlag.

DeCicco, Antonio J. & Schafer, John R. (2015). Grammatical differences between truthful and deceptive narratives. *Applied Psychology in Criminal Justice*, vol. 11(2), 75–92.

De Paulo, Bella M., Lindsay, James J., Malone, Brian E., Muhlenbruck, Laura, Charlton, Kelly & Cooper, Harris. (2003). Cues to Deception. *Psychological Bulletin*, vol. 129(1), 74–118. DOI: 10.1037/0033–2909.129.1.74.

Dietz, Simone (2003). *Die Kunst des Lügens. Eine sprachliche Fähigkeit und ihr moralischer Wert.* Hamburg: Rowohlt.

Eades, Diana (2012). The social consequences of language ideologies in courtroom cross-examination. *Language in Society*, vol. 41, 471–497. DOI: 10.1017/s0047404512000474.

Fitzpatrick, Eileen, Bachenko, Joan & Fornaciari, Tommaso (Eds.) (2015). *Automatic Detection of Verbal Deception*. DOI: 10.2200/s00656ed1v01y201507hlt029.

Fobbe, Eilika (2018). Linguistik und psychologische Täuschungsforschung – das Problem verbaler Lügenindikatoren. Unpublished talk. Conference of the *Gesellschaft für Angewandte Linguistik*, Essen (Germany), 12/9/2018.

Galasiński, Dariusz (2000). *The Language of Deception. A Discourse Analytical Study*. London/New Delhi: Sage Thousand Oaks.

Giltrow, Janet & Stein, Dieter A. (2009). *Genres in the Internet. Issues in the theory of genre.* Amsterdam: Benjamins. DOI: 10.1075/pbns.188.01gil.

Greuel, Luise (2001). *Wirklichkeit, Erinnerung, Aussage.* Weinheim: Beltz.

Günthner, Susanne (2005). Narrative reconstructions of past experiences. Adjustments and modifications in the process of recontextualizing a past experience. In Uta M. Quasthoff & Tabea Becker (Eds.). *Narrative interaction* (pp. 285–301). Amsterdam: Benjamins. DOI: 10.1075/sin.5.15gun.

Hannken-Illjes, Kati (2015). Es geht darum was Sie hier sagen. Wiedererzählen vor Gericht. In Elke Schumann, Elisabeth Gülich, Gabriele Lucius-Hoene & Stefan Pfänder (Eds.). *Wiedererzählen: Formen und Funktionen einer kulturellen Praxis* (pp. 295–316). Bielefeld: transcript. DOI: 10.14361/9783839428511–010.

Hauch, Valeria, Sporer, Siegfried L., Masip, Jaume & Blandon-Gitlin, Iris (2017). Can credibility criteria be assessed reliably? A meta-analysis of criteria-based content analysis. *Psychological Assessment*, vol. 29(6), 2017, 819–834. DOI: 10.1037/pas0000426.

Hettler, Stefanie (2006). *Wahre und falsche Zeugenaussagen*. Stuttgart: VDM Verlag Dr. Müller.

Horn, Larry R. (2017). Telling it slant: Toward a taxonomy of deception. In Janet Giltrow & Dieter Stein (Eds.), *The Pragmatic Turn in Law. Inference and Interpretation in Legal Discourse.* Mouton Series in Pragmatics (pp. 23–55). Boston: deGruyter/Mouton. DOI: 10.1515/9781501504723–002.

Labov, William, & Waletzky, Joshua (1997). Narrative analysis: Oral versions of personal experience. *Journal of Narrative & Life History*, vol. 7(1–4), 3–38. DOI: 10.1075/jnlh.7.02nar.

Meibauer, Jörg (2014). *Lying at the semantics-pragmatics interface*. Boston/Berlin: de Gruyter. DOI: 10.1515/9781614510840.

Meibauer, Jörg (2005). Lying and falsely implicating. *Journal of Pragmatics*, vol. 37, 1373–1399. DOI: 10.1016/j.pragma.2004.12.007.

Newmann, Matthew L., Pennebaker, James W., Berry, Diane S. & Richards, Jane M. (2003). Lying words: Predicting Deception from Linguistic Styles. *Personality and Social Psychology Bulletin*, vol. 29(5), 665–675. DOI: 10.1177/0146167203029005010.

Niehaus, Susanna (2008). Merkmalsorientierte Inhaltsanalyse. In Volbert, Renate & Steller, Max (Eds.), *Handbuch der Rechtspsychologie* (pp. 311–321). Göttingen: Hogrefe. Retrieved from researchgate.net/publication/260002222_Merkmalsorientierte_Inhaltsanalyse.

Norrick, Neal R. (1998). Retelling stories in spontaneous conversation. *Discourse Processes*, vol. 25(1), 75–97. DOI: 10.1080/01638539809545021.

Olsson, John (2004). *Forensic Linguistics. An Introduction to Language, Crime and the Law.* London: Continuum. DOI: 10.1558/ijslll.v11i2.298.

Quasthoff, Ulrich (1980). *Erzählen in Gesprächen. Linguistische Untersuchungen zu Strukturen und Funktionen am Beispiel einer Kommunikationsform des Alltags.* Tübingen: Narr.

Pennebaker, James W., Francis, Martha E. & Booth, Roger J. (2001). *Linguistic Inquiry and Word Count (LIWC).* Mahwah, NJ: Lawrence Erlbaum Associates.

Quijano-Sánchez, Lara, Liberatore, Frederico, Camacho-Collados, José & Camacho-Collados, Miguel (2018). Applying automatic text-based detection of deceptive language to police reports: Extracting behavioral patterns from a multi-step classification model to understand how we lie to the police. *Knowledge-Based Systems*, vol. 149, 155–168. DOI: 10.1016/j.knosys.2018.03.010.

Rickford, John R. & Sharese, King (2016). Language and linguistics on trial: hearing Rachel Jeantel (and other vernacular speakers) in the courtroom and beyond. *Language*, vol. 92/4, 948–988. DOI: 10.1353/lan.2016.0078.

Schiffrin, Deborah (2006). *In other words: Variation in reference and narrative.* Cambridge: Cambridge University Press. DOI: 10.1017/CBO9780511616273.

Steck, Peter, Hermanutz, Max, Lafrenz, Bianca, Schwind, Domenica, Hettler, Stephanie, Maier, Barbara & Geiger, Susanne. (2010). *Die psychometrische Qualität von Realkennzeichen*, 1–16. Available at core.ac.uk/download/pdf/35285699.pdf.

Stein, Dieter A. (2017). Sprachwissenschaftliche Aspekte rechtstheoretischer Ansätze im Überblick. In Ekkehard Felder & Friedemann Vogel (Eds.), *Handbuch Sprache und Recht, Handbücher Sprachwissen* (pp. 141–154). Berlin/Boston: de Gruyter. DOI: 10.1515/9783110296198–007.

Steller, Max (1989). Recent Developments in Statement Analysis. In John C. Yuille (Ed.), *Credibility assessment: proceedings of the NATO Advanced Study Institute on Credibility Assessment. Maratea, Italy, 14 - 24 June 1988* (pp. 135–154). Dordrecht: Kluwer. DOI: 10.1007/978–94–015–7856–1_8.

Steller, Max. & Köhnken, Günter (1989). Criteria-based statement analysis: Credibility assessment of children's statements in sexual abuse cases. In Jonathan D. Raskin (Ed.), *Psychological methods for investigation and evidence* (pp. 217–245). New York: Springer.

Steller, Max, Wellershaus, Petra & Wolf, Thomas (1992). Realkennzeichen in Kinderaussagen: Empirische Grundlagen der Kriterienorientierten Aussageanalyse. *Zeitschrift für Experimentelle und Angewandte Psychologie*, 39(1), 151–170.

Suchotzki, Kristina (2015). Cognitive Factors of deception, *International Journal of Psychophysiology*. vol. 95, 395–405. DOI: 10.1016/j.ijpsycho.2015.01.010.

Taylor, Paul J., Larner, Samuel, Conchie, Stacey M. & Menacere, Tarek (2017). Culture moderates changes in linguistic self-presentation and detail provision when deceiving others. *Royal Society Open Science*, vol. 2017(4), 1–20. DOI: 10.1098/rsos.170128.

Trinch, Shonna (2003). *Latina's narratives of domestic abuse: Discrepant versions of violence.* Amsterdam: Benjamins. DOI: 10.1075/impact.17.

Undeutsch, Udo (1989). The Development of Statement Reality Analysis. In John C. Yuille (Ed.), *Credibility assessment: proceedings of the NATO Advanced Study Institute on Credibility Assessment. Maratea, Italy, 14 - 24 June 1988* (pp. 101–120). Dordrecht: Kluwer. DOI: 10.1007/978–94–015–7856–1_6.

Undeutsch, Udo (1984). Courtroom evaluation of eyewitness testimony. *International Review of Applied Psychology*, vol. 33, 51–67. DOI: 10.1111/j.1464–0597.1984.tb01416.x.

Undeutsch, Udo (1967). Beurteilung der Glaubhaftigkeit von Aussagen. In Udo Undeutsch (Ed.). *Forensische Psychologie. Handbuch der Psychologie. Vol. 11* (pp. 26–181). Göttingen: Verlag für Psychologie.

Volbert, Renate & Steller, Max (2014a). Besonderheiten bei der aussagepsychologischen Begutachtung von Kindern. In Thomas Bliesener, Friedrich Lösel & Günter Köhnken (Eds.). *Lehrbuch der Rechtspsychologie* (pp. 408–421). Bern: Huber. DOI: 10.1016/b978–3-437–22902–2.00039–0.

Volbert, Renate & Steller, Max (2014b). Is this testimony truthful, fabricated, or based on false memory? Credibility assessment 25 years after Steller and Köhnken (1989). *European Psychologist*, vol. 19, 207–220. DOI: 10.1027/1016–9040/a000200.

Volbert, Renate & Steller, Max (2014c). Glaubhaftigkeit. In Thomas Bliesener & Friedrich Lösel & Günter Köhnken (Eds.). *Lehrbuch der Rechtspsychologie* (pp. 391–407). Bern: Huber. DOI: 10.1016/b978–3-437–22902–2.00039–0.

Volbert, Renate (2008). Suggestion. In Renate Volbert & Max Steller (Eds.), *Handbuch der Rechtspsychologie* (pp. 331–341). Göttingen: Hogrefe. DOI: 10.1515/juru.2009.482.

Vrij, Aldert, Fisher, Ronald & Blank, Hartmut (2017). A cognitive approach to lie detection: A meta-analysis. *Legal and Criminological Psychology*, vol. 22(1), 1–21. DOI: 10.1111/lcrp.12088.

Vrij, Aldert (2015). Verbal lie detection tools: statement validity analysis, reality monitoring, and scientific content analysis. In Pär-Anders Granhag, Aldert Vrij, & Bruno Verschuere (Eds.), *Detecting deception: Current challenges and cognitive approaches* (pp. 3–35). DOI: 10.1002/9781118510001.ch1.

Vrij, Aldert (2014). Detecting lies and deceit: Pitfalls and opportunities in nonverbal and verbal lie detection. In Charles R. Berger (Ed.), *Interpersonal Communication* (pp. 321–346). Berlin/Boston: De Gruyter. DOI: 10.1515/9783110276794.321.

Vrij, Aldert (2009). *Detecting lies and Deceit. Pitfalls and Opportunities*. Second edition. Chichester: Wiley.

Vrij, Aldert (2005). Criteria-Based Content Analysis: A Qualitative Review of the First 37 Studies. *Psychology, Public Policy, and Law*, vol. 11, 1 (3), 3–41. DOI: 10.1037/1076–8971.11.1.3.

Wegener, Hermann. (1989). The Present state of statement analysis. In John C. Yuille (Ed.), *Credibility assessment. NATO Advanced Study Institute on Credibility Assessment. Proceedings of the Conference in Maratea, Italy, 14 - 24 June 1988* (pp. 121–133). Dordrecht: Kluwer. DOI: 10.1007/978–94–015–7856–1_7.

Welle, Ida, Berclaz, Michel, Lacasa, Marie-José & Niveau, Gèrard (2016). A call to improve the validity of criterion-based content analysis (CBCA): Results from a field-based study including 60 children's statements of sexual abuse. *Journal of Forensic and Legal Medicine*, vol. 43, 111–119. DOI: 10.1016/J.Jflm.2016.08.001.

Yuille, John C. (1989). Preface. In Yuille, John C. (Ed.), *Credibility assessment. NATO Advanced Study Institute on Credibility Assessment. Proceedings of the Conference in Maratea, Italy, 14 - 24 June 1988* (pp. VII-XII). Dordrecht: Kluwer. DOI: 10.1007/978–94–015–7856–1.