HyperHamlet – Intricacies of Data Selection

Sixta Quassdorf (Basel)

Abstract

HyperHamlet is a database of allusions to and quotations from Shakespeare's *Hamlet*, which is supported by the Swiss National Science Foundation as a joint venture between the Departments of English and German Philology, and the Image & Media Lab at the University of Basel. The compilation of a corpus, whose aim it is to document the "Shakespeare phenomenon", is intricate on more than one level: the desired transdisciplinary approach between linguistics, literary and cultural studies entails data selection from a vast variety of sources; the pragmatic nature of intertextual traces, i.e. their dependence on and subordination to new contexts, further adds to formal heterogeneity. This is not only a challenge for annotation, but also for data selection. As the recognition of intertextual traces is more often than not based on intuition, this paper analyses the criteria which underlie intuition so that it can be operationalised for scholarly corpus compilation. An analogue to the pragmatic model of ostensive-inferential communication with its three constitutive parts of *speaker's meaning*, *sentence meaning* and *hearer's meaning* has been used for analytical heuristics. Authorial *intent* – in a concrete as well as in an abstract historical sense – *origin* and specific *encyclopaedic knowledge* have been found to be the basic assumptions underlying data selection, while *quantitative* factors provide supporting evidence.

1 Introduction

1.1 The Corpus

Shakespeare is generally said to have considerably contributed to the lexicon and phrase stock of the English language, yet so far the documentation of this truism has been more anecdotic than systematic. To amend this situation, a collection of quotations and allusions is being assembled at Basel University. *Hamlet*, Shakespeare's most famous drama, was chosen as an exemplary starting point. The database takes the form of a hypertext of *Hamlet* in which clickable links provide access to texts in which individual lines have been quoted since 1600. Today the *HyperHamlet* corpus comprises roughly 6,000 data sets by nearly 2,500 authors. The corpus is publicly accessible on *www.hyperhamlet.unibas.ch* for both browsing and contributing new references. Contributions from the public domain (which are edited by the project team to warrant a scholarly standard) amount to about 15 percent of the corpus. Cross-search options for parameters such as date, language, modification patterns etc. are not yet fully available to the public.

In a strict sense, *HyperHamlet* is a paradox – it is both a specialised and a reference corpus. It specialises in quotations and allusions, while admitting data from any language and period, from fiction and non-fiction, the visual arts and music, print and digital media, formal and informal settings. This broad source base was chosen with good reason:

- It enables the sister disciplines of linguistics, literary and cultural studies to join forces for the research on intertextual phenomena, which cannot be studied in a monodisciplinary way.[1]
- It documents the cultural, literary and linguistic importance of Shakespeare and meets a vast variety of research interests. In this way, the "Shakespeare phenomenon" can be studied beyond the traditional concepts of "reception studies" or "influence".
- Qualitatively, a great variety of quotative and allusive practices is covered and made available for further study.
- Quantitatively, the notable scarcity of specific expressions even in large text collections can be balanced (e.g. the phrase "the mind's eye", for which *HyperHamlet* lists 91 tokens, occurs not even once in the British National Corpus).

However advantageous, this wide range of sources entails a great heterogeneity of data, which is challenging for the corpus builders – most obviously with regard to annotation. In addition to bibliographical data, our annotation system includes categorization for

- type of reference, i.e. lexical, motif, name
- language
- year of composition
- extend of intertextual overlap, e.g. noun phrase, adjective phrase, verb phrase, clause
- modification type, e.g. substitution, omission, addition
- text genre, e.g. fiction, non-fiction
- text function, e.g. paratext, body of text
- narrative function, e.g. dialogue, neutral narrator, real author
- marking for author, work and quotation
- intertextual relationship, e.g. intertextuality, hypertextuality, metatextuality (following Genette).

These annotation features are in most cases further subcategorized so that the data can be adequately described: e.g. fiction > prose (drama/poetry) > romance (crime/fantasy/gothic/children's etc.), or paratext > stage direction (title/epigraph etc.). Data selection, however, is no less intricate. What is more, selection criteria ultimately influence the quality of data and thus the motivation of annotative features in a data-driven approach to corpus compilation.

1.2 The Data

The heterogeneity of data is also rooted in the linguistic characteristics of the intertextual traces themselves: they are pragmatic in nature so that function does not normally correlate with form. Mere reiterated words do not constitute a quotation; whoever would read "Who is there?" (*Hamlet* I i) as a Shakespearean trace without additional contextual clues? The recognition of a quotation usually depends on more than formal identity between expressions: it depends, for instance, on the interlocutor's cultural or encyclopaedic knowledge, the accessibility of the knowledge at a certain time, the communicative expectation etc. Accordingly, clear-cut static formal definitions of phenomena such as allusions and quotations would necessarily be reductive (cf. Helbig 1996 and Hohl Trillini/Quassdorf 2008b for an overview of static categorisation attempts); definitions of intertextual phenomena have to retain a certain openness and must allow for multiple meanings. Human intuition can generally cope with such fuzziness. Hence, the question of how intuition can be operationalised for the construction of *HyperHamlet* is the subject of the present article.

---

[1] I am profoundly grateful for the open-mindedness and insight which Regula Hohl Trillini brought to the cross-disciplinary discussions which have made this article possible.

2. Defining the Field in Traditional Terms

The collection of quotations and allusions in a database presupposes some sort of goal definition, which, of course, narrows the field from where data are chosen. In this sense, also our corpus is reductive. Even so, our selection criteria should allow for the documentation of the widest possible variety of intertextual phenomena, including the marginalised domains of general cultural and linguistic traces. Consequently – in contrast to former intertextual work, which preferred to focus on specific authors, literary genres or marking devices – we reversed procedures and made a single, but often quoted text our starting point. The advantage is a clearly delineated source from which to choose without *a priori* restriction of the actual data.

A database needs formal and objective criteria for data selection. The project team decided for a top-down/bottom-up procedure: a preliminary understanding of what quotations and allusions are was based on definitions found in both dictionaries and extant literature. The core notions derived from this procedure can be summarised as

- core concept of quotation: a repetitive similarity between a *Hamlet* passage and another artefact
- core concept of allusion: an associative, referential quality.

In a second step, these notions had to be validated against the data.

The crystallisation of core meanings looks like a neat distinguishing feature suitable for annotation. In practice, however, the data show that expressions are more often than not both referential as well as repetitive: they inhabit a continuum between allusions without quotations, as when Orhan Pamuk evokes *Hamlet's* revenge theme in his novel *Snow* without using Shakespeare's language (2004: 233f.), and quotations without allusions, if we do not shirk the fact that frequently-used quotations such as "there is the rub!" (cf. Glucksberg 2001: 7) can lose their association with the original text and be re-applied without evoking any association with its origin. These are extreme and extremely obvious cases; a transition point of where exactly referentiality or similarity start or end is not delimitable. Allusions and quotations are not types of expressions but properties. Subsequently, even though the traditional notions of allusion and quotation are suitable for data selection, they are not discrete enough for annotation. The differentiation between formal and conceptual repetitiveness, however, that is a materially manifest sub-criterion, can be used for annotation and leads to the categorisation of reference types (lexical, thematic and onomastic references).

3 Data Selection – Common Sense and Pragmatic Theory

3.1 Intention

The next step was to determine on what grounds decisions on repetitiveness and (associative) reference are made: the analysis of intuitively chosen data allows conclusions about the basic assumption underlying intuition, and in turn renders the selection of further data explicit and standardised.

There is little argument about the recognition of intertextual traces if they are explicitly signalled by the author of the allusion or quotation. Explicit conventionalised linguistic means comprise typographical signals such as the setting-off from the surrounding text or quotation marks as exemplified in (1) and (2). Moreover, a quotation can be indicated by naming the source, i.e. by explicit mention of author, title, prominent places or characters of the work referred to, as in examples (2) and (3). A third option are metalinguistic tags like the verb *expressed* in example (2) and the noun *quote* in (4):

(1) There is neither hope nor occasion for him "to cudgel his brains about it, he has no feeling of the business". (Coleridge: 525)[2]

(2) These explicit commands of Vatican II have been, as **Hamlet expressed** it, "more honour'd in the breach than the observance". (Davies: no pag.; ch. 4)[3]

(3) His suits were still black, but of the finest cut and quality. "With a star and ribbon, and his stocking down, and his hair over his shoulder, he would make a pretty **Hamlet**", said the gay old Duchess Queensberry, "and I make no doubt he has been the death of a dozen **Ophelias** already, here and amongst the Indians", she added, thinking not at all the worse of Harry for his supposed successes among the fair. (Thackeray: 246f.)[4]

(4) The following morning, Peter Gilbert, still in his pyjamas, intruded upon me in the bathroom while I was shaving. "It's customary to knock", I said. "A custom more honoured in the breach than the observance?" he said, and inflected it upwards to indicate that he considered **the quote** so apt as to be witty. (Gott: 17)

These signals overtly indicate that some part of the message can only be recovered if another source is taken into account. Yet they do not necessarily point to a concrete (literary) source, they merely direct "the audience to some item association with *X* other than its extension" (Saka 1998: 126). The actual source may remain obscure, unless the reader can link the quoted elements to her encyclopaedic knowledge. Even if in (3) the name *Hamlet* is not recognised as that of a famous play or character, the addressee will infer a symbolic function. In all four examples, the writer provides the essential "meta-allusive" information which necessarily presupposes an intentional act.

Communicative intent is vital in pragmatic theory. Grice claims that meaning in communication evolves because the hearer can infer the communicative intent of the speaker (*speaker's meaning*) thanks to the ostensive stimuli provided by her linguistic and other communicative behaviour (cf. Grice 1957). In *Relevance Theory*, any extra-effort taken, e.g. the overt hinting at an additional layer of meaning is assumed to be motivated by relevance (cf. Sperber/Wilson 2007: 118–123). It is the hearer's task to infer an appropriate implicature. Accordingly, at the heart of meaning construction lies the notion of a deliberate communicative intent, which is ultimately also the basic assumption underlying the recognition of quotative and allusive data. Sperber/Wilson even go so far as to declare that the "attribution of intentions to others is a characteristic feature of human cognition and interaction" (2007: 23f.).

Stylistic anomalies such as archaic forms in (5) or passages in a different language as in (6) also induce the addressee to infer the communicative intent of conveying more than just the *sentence meaning* (cf. Grice 1957). Again, the effort of changing the linguistic register is assumed to be motivated. The addressee will presuppose relevance and infer further implicatures, i.e. that a more or less fixed expression has been used (cf. Sperber/Wilson 2007: 158):

(5) Wexford said thoughtfully, "He **doth** protest too much." "About the Kingtons' mutual devotion, do you mean?" asked Burden. "That was a strange remark." (Rendell:104)[5]

(6) Das Volk in seinen Urwahlen besässe die *Freiheit der äusseren Bewegung*. Aber die *innere Freiheit*? **That is the question**! (Marx: 28)[6]

---

[2] A conflation of two Hamlet passages: "Cudgel thy brains no more about it" (V i 47) and "Has this fellow no feeling of the business?" (V i 55). Explicit signals are highlighted in bold print.

[3] Cf. "it is a custom / More honour'd in the breach than the observance." (I iv).

[4] Cf. also: "No hat upon his head; his stockings foul'd, / Ungarter'd, and down-gyved to his ancle"; Onomastic marking is here used to indicate a thematic reference to Hamlet's appearance as described in Ophelia's report (II i).

[5] Cf. "The lady protests too much, methinks." (III ii).

[6] Cf. "To be, or not to be: That is the question." (III i). The italics are original.

For completeness' sake, also *genre* has to be mentioned in this section: dictionary collections of quotations, allusions and famous texts, as well as epigrams contain quotations by definition.

The above mentioned criteria are discrete enough to work for straightforward intersubjective identification and for annotation. Needless to say, these signals often occur in combination, which strengthen the assumption of prior conscious intertextual intent. Our data as well as the requirements posed by our objective of systematisation for a searchable database have lead to the conclusion that marking devices can objectively be distinguished by identifiable overt linguistic marking strategies which, in turn, are attributable to prior authorial intent. It is noteworthy, that these marking devices can be used to produce pseudo-quotations, i.e. they convey a communicative intent which is ultimately independent from the origin of the marked passage itself.[7]

3.2 Origin

Language users, however, may prefer not to make their intention fully manifest. The choice of explicitness or implicitness depends on different subjective and objective preconditions such as assumed shared knowledge or a writer's inclination for word play. Furthermore, the often-discussed function of *mention* may be so secondary and the function of *use* (cf. Saka 1998) so predominant to the communicative intent that source indications would communicatively be misleading. Explicit marking is a sufficient but not a necessary expression of authorial intent, which, on the other hand, puts any reading audience and especially researchers aiming to assemble a database of intersubjectively recognisable allusions in an awkward situation: how can we be justified in assuming authorial intent without explicit signals?

In structuralism, the problem of uncovering authorial intent is deferred by relying on the semiotic code model where a signifier is linked to its signified. In the case of allusions and quotations we can analogise that the signifier is the quoting passage and the signified the link to the original text. Moreover, phraseological and corpuslinguistic investigations have shown that people tend to use similar forms to express similar meanings. The commonsense notion that quotations should be verbatim and the inversion of the argument that (near-)verbatim renditions are normally equalled with (intended) quotations can be deduced from the semiotic code model. Form may indeed be indicative of communicative intent. However, form may also reveal non-intentional information: an accent can reveal a speaker's geographical origins and a particular choice of words her previous experiences with language. Hence structuralism is able to deflect attention away from authorial intent, and instead raises questions of *origin*, which is the second basic assumption underlying intuitional choices. In literary theory, the opinion that meaning resides in the text alone and that the author is actually dead (cf. Barthes 1993) echo these structuralist insights.

Grice's notion of *natural* and *non-natural meaning* illuminates the relationship between intent and origin from another angle (cf. Grice 1957): words can be "naturally telling" about their origin. They are merely an indexical sign, a form which is open to direct perception (cf. Millikan 1998) just as a photograph lying about is a *natural* sign of a certain depicted situation. Linguistic expressions can therefore be assigned a *natural* referential quality, which operates, among other things, as an implicit stimulus for allusive or quotational interpretation. We call it *implicit* because a) there is always a possibility of chance identity/similarity and b) the quotational potential remains dormant unless the origin of the phrase is known by the addressee.

---

[7] We believe that the sophisticated discussion on marking theory which is expounded in Helbig (1996) can be simplified significantly if the complexities involved in intertextual reference are disentangled. One step towards this goal is to define marking on the grounds of conventionalised pragmatic inference, as suggested in this article. The work on annotating the HyperHamlet corpus has justified this practice.

Examples (7)–(9) illustrate some "naturally telling" similarities to the *Hamlet* text: (7) is a verbatim rendition, (8) is the translation into Modern English of the beginning of the "Closet Scene" in act III and (9) demonstrates the use of conspicuous combinations such as "alas poor" together with a longer string that is structurally and phonetically similar to the original. The origin of (9) is sufficiently identifiable despite lexical additions and substitutions:

(7) Some information is fascinating, some humbling. **What a piece of work is man** – and yet 50 percent of our genes are the same as a banana's. (Graeber: no pag.)[8]

(8) [Countess:] Lord Brookside, **you have offended your father**.
[Brookside:] **Mother, you have offended my father**. (Kubrick: no pag.)[9]

(9) "Any pain?" I asked hopefully. "Naw... they goat drugs... jist ma breathin..." I held his hand and felt a twinge of amusement as his pathetic, bony fingers squeezed tightly. I thought I was going to laugh in his skeletal face as his tired eyes kept shutting. **Alas poor Alan, I knew him Nurse. He was a wanker, an infinite pest**. I watched, stifling smirks, as he groped for breath. (Welsh: 247)[10]

The authors of (7)–(9) refrain from overt marking. Nevertheless, once the elements from *Hamlet* are recognised, the reader can be confident that the similarities are not coincidental and fall back on *intention*. Length, i.e. a quantitative notion, can strengthen the manifestness of evidence for communicative intent for at least two reasons: in the terms of *Relevance Theory*, the effort of repeating a certain passage, be it verbatim or modified, increases with length and is therefore more likely to be motivated. In a structuralist view, which makes use of probabilistic rules, the longer the passage, the less likely are chance similarities.

Another commonsense notion of "naturally telling" intertextual phenomena are coinages – the prime example for the basic notion of *origin*: what Shakespeare created is necessarily a Shakespearean trace. The trace is self-referentially "telling" and documents intertextual processes independent of both communicative intent and inferential interpretation. Yet this is a very theoretical stance! As soon as we want to ascertain what a coinage is, the question becomes vexed: the concept of *coinage* is difficult to operationalise for two reasons: a) because linguistic data from the Elizabethan period are limited and b) because not only original material is quoted.

By differentiating between types of creativity some certainty of Shakespearean originality can be assumed, again on quantitative grounds. The true origin of single items, be it a single new idea, a single new word or word form (cf. table 1: types i, v and vi) can only speculatively be traced back as we know that small items tend to be "spontaneously" formed, understood, learned and taken up if they fulfil a certain communicative purpose well. Their first appearance on paper may considerably postdate their creation. More confidence can be placed in thematic or verbal elaboration (types ii–iv) as the increase of length and complexity strengthens the belief in originality and weakens the option of chance composition.

[8] Cf. "What a piece of work is a man! how noble in reason!" (II ii).

[9] Cf. "[Gertrude:] Hamlet, thou hast thy father much offended. / [Hamlet:] Mother, you have my father much offended." (III iv).

[10] Cf. "Alas, poor Yorick! I knew him, Horatio: a fellow of infinite jest." (V i).

i.	new concepts or ideas
ii.	thematic elaboration of known concepts or ideas, e.g. Hamlet's characterisation in comparison to the template of Saxo Grammaticus' *Gesta Danorum*
iii.	new metaphors for common ideas or concepts, e.g. "to be hoist with one's own petard" (III iv), which is a new version of the proverb "the fowler is caught in his own net"
iv.	verbal elaboration of sayings, clichés or proverbs, e.g. "A little more than kin and less than kind" (I ii) derives from the proverb "The nearer in kin the less in kindness."
v.	new words and collocations which denote a particular quality, an event, a process or an object, e.g. "self-slaughter" (I ii), "primrose path" (I iii)
vi.	word formation, e.g. "buzzer" (IV v), "the avouch" (I i), "to sickly" (III i), "unanel'd" (I v).

Table 1: Types of coinages/creativity

Apart from the fact that the earliest written record of a word or phrase does not necessarily indicate its date of creation, people do not only re-apply new words and phrases. Expressions can have become popular through *Hamlet* although they were already around – a case in point may be Hamlet's famous "to be or not to be" disjunction (cf. Hohl Trillini 2009).

3.3 Encyclopaedic Knowledge

The pragmatic model comprises: firstly, a speaker who intends to communicate a certain intended meaning by more ore less ostensive communicative behaviour. This aspect of the model was used for the interpretation and delineation of marking strategies in 3.1 above. Secondly, section 3.2. discussed the *sentence meaning*, i.e. the code. Thirdly, there is the hearer who wants to recover communicative intent by inference from the clues accessible to her at a certain time in a certain situation. Nerlich/Clarke recognise even more than Grice or Sperber/Wilson the role of the addressee in meaning construction and introduce the notion of *hearer's meaning* (2001: 10). The receptive side is thus institutionalised as a constitutional part of the model. Especially in the written mode *hearer's meaning* (i.e. its analogue *reader's meaning*) is justified due to the "communicative difference" (Plett 1975: 80) between the production and the reception of a text.

Even though the commonsensical reader may generally try to recover the author's meaning, she cannot but hypothesize about communicative intent since the author is not normally available for verification. Consequently, it is a necessity rather than a convenience to allow for further, available clues which interact in meaning construction, even if they are less objectively reliable. Apart from the clues discussed above, i.e. explicit marking by the author and the recognition of "naturally telling" passages, recurrence to a broader encyclopaedic knowledge is essential. While reading literature, the reader may expect that a text contains allusions to other texts because of her general knowledge of how literature works (Meyer 1961: 22). She may pay attention to possible allusions and see it as part of her reading task to detect them (while she would read a manual with a very different bias!). The text genre can thus be seen as one of the constraints which guide readerly meaning construction.

Literary language is language at its fullest, and not seldom stretches the boundaries (cf. Coseriu 1994: 160). As a result, literary criticism has specialised in inferential interpreting processes from which linguists can profit. The subtler literary allusions are, the more they are prized by literary critics because the activation of the reader's inferential powers allows for multiplicities and complexities of potential meaning, which in turn enrich the reading experience. Knowledge about authors is a thus a further clue: several authors use references to another text not only in one single passage but have a general predilection for a source text which resurfaces throughout their entire oeuvre. Charles Dickens, amongst others, repeatedly refers

to *Hamlet*. He obviously knew "his" *Hamlet* so well that no similarity to the source text will have unintentionally escaped his pen. Hence, certain authors are "as good as a marker" for an informed reader.

Based on this experience, allusive passages brought recognition of another formal allusive strategy – clustering. In addition to the passage openly marked by quotation marks in (10), there are two more echoes of *Hamlet* which are not clearly signalled. They are fragments but can be discerned as further allusive keywords in the context of this short passage. Example (11) illustrates the same mechanism: knowing that Walter Scott is a great *Hamlet* quoter and knowing that authors often quote more than once, i.e. that they cluster, one can assume that the mention of stirring mice, rats, conscience and the allusion to suicide "by one bold stroke" – all figuring in prominent places of the Shakespearean play – are not coincidental. This constructed allusive meaning is ultimately only a *reader's meaning*, yet the clues the reader draws on follow a pattern:

(10) Do **thy prophetic Fears** anticipate, / Meek Child of Misery! Thy future fate? – / The starving meal, and **all the thousand aches** / "**Which patient Merit of th' Unworthy takes?** (Coleridge: 147)[11]

(11) "What am I now," he said to himself, "that I am thus jaded by the words of a mean, weather-beaten, goose-brained gull! **Conscience**, thou art a blood-hound, whose growl wakes as readily at the paltry **stir of a rat or mouse**, as at the step of a lion. **Can I not quit myself, by one bold stroke**, of a state so irksome, so unhonoured? What if I kneel to Elizabeth, and, owning the whole, throw myself on her mercy?" (Scott: 321f.)[12]

More generally, also knowledge about style, quoting habits of social groups, historical periods and art forms informs us about the likelihood of intertextual phenomena. Familiarity with the referential strategies of postmodernist writers or the fashion of quoting from the classics among the educated classes of the 19th c. (cf. also quotation dictionaries and anthologies which list passages that had a certain currency as a quotation at a certain period of time) affect the decision of whether or not a linguistic element is recognised as an element from *Hamlet*. General or specific encyclopaedic knowledge as well as the cotext can turn inconspicuous "normal" English words or common ideas into Shakespearean traces.

What is more, the notion of *hearer's meaning* offers an additional solution to the problem of unconscious quotation: presupposing a usage-based model of language acquisition, we may generalise from the above-mentioned processes and take the step from the token to the type. The more often an item is (intentionally) quoted, the more likely it is to become an "anonymous" expression and possibly in a later step, a lexicalised item of the *langue*. If, for example, the verb phrase "hair stand on end" is repeatedly found in explicitly or implicitly marked text passages and/or in quotation dictionaries, and no earlier trace is found,[13] then also unmarked occurrences of "hair stand on end" can be considered to be traces of the play, even though the collocation appears to be completely unobtrusive and the link to *Hamlet* utterly lost. The notion of authorial intent can now be understood in a more abstract sense: it need not apply to every concrete data set, but can be derived from a diachronic point of view.

---

[11] Cf. "O my prophetic soul! My uncle!" (I v).
"The heart-ache and the thousand natural shocks / That flesh is heir to."
"the spurns / That patient merit of th' unworthy takes." (both III i).

[12] Cf. "Thus conscience does make cowards of us all" (III i).
"Not a mouse stirring;" (I i).
"How now! a rat? Dead, for a ducat, dead! and Hamlet IV i: A rat, a rat!" (III iv).
"Or that the Everlasting had not fix'd / His canon 'gainst self-slaughter! O God! God!" (I ii).
"When he himself might his quietus make / With a bare bodkin?" (III i).

[13] Contemporaneous Bibles use the expression "hair stands up" or "upright".

4 Conclusion

The compilation of a corpus to document the "Shakespeare phenomenon" is intricate on more than one level: the desirable transdisciplinary approach entails data selection from a vast variety of sources; the pragmatic nature of intertextual traces further adds to formal heterogeneity, which is not only a challenge for annotation, but also for data selection. As the recognition of intertextual traces is more often than not based on intuition, this paper set out to analyse the criteria which underlie intuition so that it can be operationalised for scholarly corpus compilation. The pragmatic model of ostensive-inferential communication with its three constitutive parts of *speaker's meaning*, *sentence meaning* and *hearer's meaning* has been used for analytical heuristics. Authorial *intent* – in a concrete as well as in an abstract historical sense – *origin* and specific *encyclopaedic knowledge* have been found to be the basic assumptions underlying data selection, while *quantity* provides supporting evidence. According to the tripartite model, data can be grouped into:

- marked or overt quotations and allusions requiring general knowledge of linguistic conventions;
- implicit or "naturally telling" quotations requiring knowledge of the source text together with general linguistic knowledge, and
- covert traces requiring general or expert encyclopaedic knowledge in interaction with knowledge of the source text and general linguistic knowledge.

These differentiations are, however, not necessarily suitable for data annotation. Whereas the characteristics of overtly marked quotations and allusions are discrete and can thus be directly used for annotation, implicit and covert references form a continuum. More fine-grained features had to be identified: the answers to questions such as "what is referred to?", "where does a reference occur?", "who made the reference?", "when was a reference made?", "how is a reference integrated?", "is the reference modified?" etc. underlie data selection in a more specific manner and provide the framework for data annotation.

Altogether, these linguistic and encyclopaedic clues serve to identify Shakespearean traces despite formal variability, fragmentation and lexicalisation. They also establish a filter for phrases which did not need Shakespeare to make their way into present day English (such as *Who is there?*). As a result, our empirical hermeneutic approach has taken us one step further in intertextual theory, the theory of meaning construction and the methodology of managing heterogeneous data without *a priori* restrictions.

References

Primary literature

Coleridge, Samuel Taylor (1995): *The Collected Works of Samuel Taylor Coleridge*. 16 vols. Vol. 11/1: *Theory of Life*. Princeton, Princeton University Press: 481–557.

Coleridge, Samuel Taylor (2001): *The Collected Works of Samuel Taylor Coleridge*. 16 vols. Vol. 16/1/1: *To a Young Ass, Its Mother Being Tethered near It*. Princeton, Princeton University Press: 146–148.

Davies, Michael (1995): *Liturgical Shipwreck: 25 Years of the New Mass 1969–1994*. Rockford, IL: Tan Books.

Glucksberg, Sam (2001): *Understanding Figurative Language: From Metaphors to Idioms*. Oxford: Oxford University Press.

Gott, Robert (2007): *Amongst the Dead*. Melbourne: Scribe.

Graeber, Laurel (2007): "Spare Times: 'Genome: The Secret of How Life Works'". *The New York Times*, 6 April 2007.

Kubrick, Stanley (1975): *The Memoirs of Barry Lyndon, Esq*. Hollywood: Warner Bros.

Marx, Karl (1959): "Bekenntnisse einer schönen Seele". *Neue Rheinische Zeitung*. 17 November 1848. *Marx, Karl/Engels, Friedrich: Werke*. 43 vols. Vol. 6. Berlin, Dietz Verlag: 24–28.
Pamuk, Orhan (2004): *Snow*. New York: Vintage.
Rendell, Ruth (1988): *The Speaker of Mandarin: An Inspector Wexford Mystery*. London: Hutchinson.
Scott, Sir Walter (1993): *Kenilworth: A Romance*. *The Edinburgh Edition of the Waverley Novels*. 30 vols. Vol. 11. Edinburgh, Edinburgh University Press.
Thackeray, William Makepeace (1899): "The Virginians: A Tale of the Last Century". *The Works of William Makepeace Thackeray*. 13 vols. Vol. 10. London, Smith, Elder & Co.: 246f.
Welsh, Irvine (1996): *Trainspotting*. London: Minerva.
Saxo Grammaticus (1979): "Amleth, Prince of Denmark". *Gesta Danorum*: Books I–IX. 2 vols. Cambridge: Brewer.

Secondary literature

Barthes, Roland (1993): "The Death of the Author". In: Lodge, David (ed.): *Modern Criticism and Theory: A Reader*. London and New York, Longman: 167–172.
Clark, Herbert H./Wade, Elizabeth (1993): "Reproduction and Demonstration in Quotations". *Journal of Memory and Language* 32/6: 805–819.
Coseriu, Eugenio (1994): *Textlinguistik*. Tübingen and Basel: Francke.
Genette, Gérard (1997): *Palimpsests: Literature in the Second Degree*. Lincoln: University of Nebraska Press.
Grice, Herbert Paul (1957): "Meaning". *Philosophical Review* 66: 377–388.
Grice, Herbert Paul (1975): "Logic and Conversation". In: Cole, Peter/Morgan, Jerry L. (eds.): *Syntax and Semantics*. Vol. 3. New York, Academic Press: 41–58.
Helbig, Jörg (1996): *Intertextualität und Markierung: Untersuchungen zur Systematik und Funktion der Signalisierung von Intertextualität*. Heidelberg: Winter.
Hohl Trillini, Regula (2009): "*Hamlet* and Textual Re-Production: The Case of 'To Be or Not to Be' (1561–1726)". *Swiss Papers in Literature and Language* 22 (forthcoming).
Hohl Trillini, Regula/Quassdorf, Sixta (2008a): "Quotations and their Co(n)Texts". In: Hamm, Albert/Higgs, Lyndon (eds.): *Variability and Change in Language and Discourse*. Strasbourg, Université Marc Bloch: 77–89.
Hohl Trillini, Regula/Quassdorf, Sixta (2008b): "A 'Key to all Quotations'? A Corpus-Based Parameter Model of Intertextuality". Submitted manuscript.
Lennon, Paul (2004): *Allusions in the Press: An Applied Linguistic Study*. Berlin, New York: Mouton de Gruyter.
Meyer, Herman (1961): *Das Zitat in der Erzählkunst. Zur Geschichte und Poetik des europäischen Romans*. Stuttgart: Metzler.
Millikan, Ruth (1998): "A More Plausible Kind of 'Recognitional Concept'". In: Villanueva, Enrique (ed.): *Concepts*: *Philosophical Issues*. Atascadero, Ridgeview Publishing: Vol. 9, 35–41.
Nerlich, Brigitte/Clarke, David D. (2001): "Ambiguities We Live By: Towards a Pragmatics of Polysemy". *Journal of Pragmatics* 33: 1–20.
Plett, Heinrich F. (1975): *Textwissenschaft und Textanalyse*. Heidelberg: Quelle & Meyer.
Plett, Heinrich F. (1988): "The Poetics of Quotation". *Annales Universitatis Scientiarium Budapestiensis*. Sectio Linguistica 17: 293–313.
Saka, Paul (1998): "Quotation and the Use-Mention Distinction". *Mind* 107/425: 113–135.
Sperber, Dan/Wilson, Deirdre (2007): *Relevance: Communication and Cognition*. Oxford: Blackwell.
Sternberg, Meir (1982): "Proteus in Quotation Land". *Poetics Today* 3/2: 107–156.

Tuomarla, Ulla (2000): *La citation mode d'emploi: Sur le fonctionnement discursif du discours rapporte direct*. Helsinki: Academia Scientiarum Fennica.
Wray, Alison (2002): *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.

ISSN 1615-3014