ASPECTS OF TEXT MININGFROM COMPUTATIONAL SEMIOTICS TO SYSTEMIC FUNCTIONAL HYPERTEXTS

Alexander Mehler
University of Trier
mehler@uni-trier.de

ABSTRACT

The significance of natural language texts as the prime information structure for the management and dissemination of knowledge in organisations is still increasing. Making relevant documents available depending on varying tasks in different contexts is of primary importance for any efficient task completion. Implementing this demand requires the content based processing of texts, which enables to reconstruct or, if necessary, to explore the relationship of task, context and document. Text mining is a technology that is suitable for solving problems of this kind. In the following, semiotic aspects of text mining are investigated. Based on the primary object of text mining – *natural language texts* – the specific complexity of this class of signs is outlined and requirements for the implementation of text mining procedures are derived. This is done with reference to *text linkage* introduced as a special task in text mining. Text linkage refers to the exploration of implicit, content based relations of texts (and their annotation as typed links in corpora possibly organised as hypertexts). In this context, the term *systemic functional hypertext* is introduced, which distinguishes genre and register layers for the management of links in a poly-level hypertext system.

INTRODUCTION

The set of documents available online increases rapidly. This is evident when regarding the availability of text, image, audio and video documents as part of intranets, extranets or the Internet. The majority of these documents are *natural language texts* lacking schematic structures resulting from the usage of forms or text building blocks.[2] Due to the enormous increase of documents for the textual representation of business and public organisations, an immense problem of information processing emerges, which is tied to the following exemplary questions: Which documents are relevant for which task in which context? How does an organisation's member find these documents? Has any document been omitted, and if so, in which sense? These questions cannot be answered solely with the help of information retrieval techniques. Information retrieval is concerned with the problem of retrieving indexed texts from pre-processed document collections on the basis of structured queries as part of standardised retrieval languages.[3]

In contrast to this, the problem of information processing outlined above is characterised by its *uncertainty* – or more specifically: by its vagueness and ambiguity[4] – as well as by its *contextsensitivity* and *dynamics*. Among other aspects, *vagueness* is a type of uncertainty referring to the impossibility of making precise distinctions and assignments, whereas *ambiguity* denotes the existence of diverse, dissonant, nonspecific, and possibly competing alternatives. With regard to the information processing, these types of uncertainty can be characterised as follows: Given a task T connected with a possibly vague information need I in context C, it may not be categorically decidable, whether a document x is relevant for T depending on C, or not (contextsensitivity and vagueness).[5] In addition, there are possibly many documents that are assumed to be relevant for T in context C (ambiguity in the sense of nonspecificity). Since there may exist a collection of texts, each of which cannot be categorically assigned to the given task, vagueness and ambiguity are types of uncertainty, which – besides context sensitivity – simultaneously apply to the task of text mining. Beyond context sensitivity and uncertainty, the *dynamics* of the relation of task, context, and text has to be considered: as organisations, their tasks and contexts change, the evaluation of documents as being relevant changes, too. To summarise: there is not only a fuzzy many to many relation between texts, contexts, and tasks (the same text can be relevant for the same task to varying degrees in different contexts), but this relation has a further parameter: time. Due to the existence of

[2] Clearly, this "lack" does not indicate a deficient mode of textual organization. On the contrary, natural language texts represent the most efficient information structure for the dissemination of knowledge in human communities.

[3] See Sparck Jones (1999:7).

[4] See Zadeh (1997) for a constraint-based formalization of different terms of uncertainty. See also Klir/Folger (1988) for information theoretical definitions of these terms.

[5] The term *task* – as well as the term *organization* – is used in the following rather unspecifically. A task is a generic term for any target, which demands a specific state to be realised by members of a given organization. In case of economic organizations, the general operating purpose of the corresponding enterprise (determined by the production of certain goods or the provision of services) determines its tasks. For a text to be relevant for performing a specific task means for example that processing this text may help to reach the goal defined by the task. In a more elaborate model these organization theoretical terms and their relations need to be defined in more detail.

varying contexts for judging a text as task relevant, which cannot be explicitly coded in the document collection under consideration, an information processing problem emerges, whose degree of virulence correlates with the grade of uncertainty of the dynamic mapping of texts, contexts, and tasks. In any case, a risk to overlook relevant or to receive irrelevant documents emerges.[6] Now, three problem categories of increasing complexity can be distinguished:[7]

- *The retrieval problem:* a known already structured information space S is confronted with determinate, structured information needs. This problem category – characterising the classical field of application of information retrieval and database systems – is associated with questions of the following type: *Which information units of the already structured information space S have the features* $F_1,\square,F_n$*?* In information retrieval, features are index terms and S is a collection of indexed documents. In the area of database systems, features are constituents of a database query, and S is a relational or object oriented database. Though the mapping of information units and features F_i, $i \in \{1,\ldots,n\}$, may be uncertain (as in case of fuzzy databases or graded relevance judgements of retrieved documents), the set of information units and their features is supposed to be determined. Contexts can only be modelled by means of enumerable, static features F_i. The retrieval problem presupposes the solution of categorisation tasks:
- *The categorisation problem:* an unknown, unstructured information space S is confronted with determinate structured information needs.[8] This scenario is characterised by questions of the following type: *To which features* $F_1,\square,F_n$ *do the information units of S belong?* This question goes beyond the domain of the retrieval problem since it does not presuppose that the information units are part of an already structured information space. Though the set of features F_i is supposed to be known (e.g. as being elements of a thesaurus, a set of index terms, or a set of topic categories), it is unknown, which information units have which features. The categorisation problem is addressed in the area of text classification. It clearly arises during the automatic build-up and maintenance of IR systems. From the perspective of an organisation's member, a categorisation problem emerges when she is able to describe her information need, but does not know how to connect her description with the unknown information space.
- *The exploration problem:* an unknown, unstructured, dynamic information space is confronted with an uncertain, unstructured information need. This problem category is connected with the characteristic question: *Which information units are or become relevant for which task in which sense under which circumstances?* Regarding the categorisation problem, the presupposition is abandoned that the relevant features describing the information need are known, static, context insensitive, and unstructured. From a member's perspective, this problem arises when she is not conscious of her information need and does not know much or even anything about the documents available, their contents and respects that make them relevant for her task. This is the kind of information processing problem described above with recourse to the terms of uncertainty, contextsensitivity, and dynamics.

These three interdependent problem categories can be illustrated as follows: Suppose, a member of an organisation deals with a task T_i described by a set of directly available documents X_i as part of the information space S. If the organisation disposes of an information retrieval system and the member knows to describe her information need based on the corresponding query language, she can formulate a query in order to select task relevant documents $X \square X_i$. Hence, a retrieval problem arises.[9] If there is no retrieval system or if it only covers a subset of S, the content based connection of T_i and S has to be detected, first. Suppose that there exists knowledge about characteristic features of T_i, e.g. with respect to the type to which it belongs. In this case, a categorisation problem arises: those documents, which are categorised by the known features of T_i, have to be determined. In the case of missing knowledge, an exploration problem arises that can be solved approximately as follows: Maybe there exists a task T_j (not necessarily known to the user), which is connected with a document set X_j (for the description of T_j and its realisation) and is somehow related to task T_i (maybe, T_i is a sub-, subsequent or concurrent task of T_j, etc.). In this case the (type of) relation of T_i and T_j has to be discovered and documents serving for describing both complexes have to be linked. Since the document set X_i is supposed to be the only available information-describing task T_i, it serves (among others) as an informational basis for exploring such

[6] Users do not survey the documents available, their contents and the respects under which they may become relevant for their tasks (information needs).

[7] These three categories only describe extreme cases of a continuum of more subtle tasks.

[8] The case of a structured information space (e.g. an indexed document collection of an information retrieval system) which is confronted with an unstructured information need will not be taken into consideration. It may be seen as a *learning problem* of how to use the corresponding query language in order to select information units from the information space.

[9] In this context, the property of the information space to be known means that the documents which the member is seeking are supposed to be an already indexed part of I, so that they can be requested using the retrieval language.

content based, text-spanning relations. This is illustrated in figure (1) by means of a commuting diagram, where r denotes the relevance relation between tasks and documents, □ is the type of dependence of T_i and T_j, and ↔ parallels this dependence on the level of documents. On the background of the dependence of T_i and T_j it is explored that the set of documents X_j is related to T_i. An alternative, bottom-up strategy would be to explore X_i in order to derive features describing T_i and then to look for documents in S, which are fuzzily categorised by these features. As tasks T_i, T_j change, this procedure needs to be repeated.

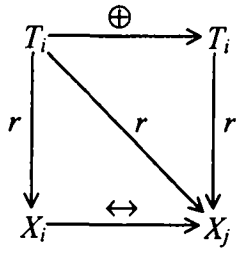

Figure 1: The (intertextual) dependence of documents on the background of two tasks.

The solution of the retrieval problem presupposes that all items which can be retrieved are determined by the indexation of the information space S. Items that are not indexed cannot be retrieved. The retrieval capacity is determined by the antecedent process of indexation, and does not include processes of reconstruction or even exploration. In contrast to this, the exploration problem arises from the twofold uncertainty, contextsensitivity and dynamics of information need and information space: not only does the contextsensitive connection between information need and information units have to be explored, but it also varies as the organisational context evolves. In other words, the exploration problem cannot be solved by indexation techniques, because the universe of (social-semiotic) contexts cannot be extensionally represented ex ante. The categorisation problem has an exceptional position in the sense that a solution of the exploration problem includes a solution of the retrieval problem, which on its part requires a solution of the categorisation problem.
The automatic solution of these three tasks primarily demands the use of *minimal knowledge resources:* instead of presupposing a broad range of knowledge derived from human expertise, a sparse, procedural model is required, which is able to *learn* relevant knowledge from natural language texts, task or context descriptions. This is a prerequisite for the re-usability, adaptability, and transferability of the same solution in an area of rapidly changing tasks, contexts, and domains of interest.

TEXT MINING

Text mining is a technology that is especially suitable for solving problems of the type of the exploration task outlined above. Objects of text mining are natural language texts, which lack the explicit, schematic organisation of formal data structures. Instead of providing a nominal definition of text mining, it is operationally outlined by specifying the order of typical text mining procedures.[10] Given any text corpus C, the following text mining steps can be distinguished:

1. **Reconstruction:** it is the *reconstruction* of linguistic knowledge underlying the usage of text components and their intra-, intertextual and exophoric relations that occurs first. "Reconstruction" means that knowledge is *discovered* – and *not* simply matched with predefined schemata or patterns – that was (or has been) constitutive for the production (reception) of texts in C.[11] Instead of starting with predefined, fixed and restricted knowledge domains, reconstruction aims at replacing *factual* by *procedural* knowledge of how to acquire/learn (domain or task) relevant knowledge. This processing stage is not bound to the analysis of corpus C, but may include the processing of a (possibly much larger) training corpus. For this purpose, methods of computational linguistics, machine learning, knowledge engineering (data mining), statistics and many other fields are applied.
2. **Systematisation**: secondly, the *systematisation* of the structures reconstructed in the previous stage occurs. One has to examine, which *types* (classes or clusters) are to be distinguished based on which kind of text function. A type allows to subsume linguistic units showing similar or homogeneous usage regularities

[10] See Hearst (1999) for a "negative" definition of what she calls *text data mining*, which she distinguishes from information retrieval, data mining, and corpus-based computational linguistics.

[11] See also Hearst (1999).

with respect to a given function (in the corpus under consideration). Systematisation aims at reconstructing higher level linguistic or conceptual units, such as types of rhetorical structure, schematic structure, etc. Reconstruction and systematisation are not strictly ordered. They may be organised with the help of a feedback mechanism so that relations of types (classes) are reconstructed on their own, which are used for evaluating (confirming/modifying) representations of relations of their instances.

3. **Exploration:** subsequently, heretofore unknown, task relevant, structural or referential relations, which were *not* (necessarily) constitutive for the texts under consideration, are *explored.*[12] This can be exemplified as follows: suppose a situation, in which *n* persons produce texts, each of which is only known by the corresponding author. Tthere may be relationships of semantic similarity helping to facilitate, clarify or augment the reception of these texts so that the reader gains information which she cannot obtain from processing any of these texts in isolation. As in case of the distinction of categorisation and exploration tasks, reconstruction, systematisation and exploration are not categorically separated.

4. **Extraction,** *annotation and visualisation*: finally, implicit relations that were discovered during the first three steps are made explicit with the help of annotation and visualisation techniques. This step has the aims of making *explicit* the information reconstructed or explored before so that a user can easily skim the value-added information inherent to the texts under consideration. Examples are the annotation of attributes (and their values) as well as relations (links) of text components (using a standardised annotation language like XML), their visualisation in hierarchies or networks, the summarisation of single texts, or the response to natural language queries regarding a text base.

Obviously, there cannot be a knowledge free discovery approach. Rather, text mining relies on sparse, transparent, automatically adaptable and extendable resources. Its *procedural knowledge* applied during the reconstruction stage primarily aims at supplying the subsequent mining procedures with cotextual and contextual knowledge derived from a natural language corpus of texts without the need to refer to introspectively predefined knowledge.

Text mining should not be confused with *data mining*. Data mining is a technology that aims to extract relevant, previously unknown information from relational or object oriented databases.[13] In contrast to this, text mining aims at exploring relevant, previously unknown knowledge from natural language texts, which are characterised by their implicit, indeterminate organisation which makes reconstruction an indispensable part of text mining. Text mining can be seen as covering the following tasks:[14]

- **Text knowledge engineering** aims at the maintenance and acquisition of domain-specific knowledge (organised as conceptual taxonomies) from natural language texts. Hahn/ Schnattinger (1998) explicitly separate text knowledge engineering from *information extraction*: in contrast to the latter, the former does not aim at simply instantiating predefined templates with factual knowledge, but at automatically learning and enhancing conceptual, taxonomically organised structures based on (partial) processing of natural language texts.[15]
- **Text categorisation** aims at classifying texts by mapping them to a set of predefined categories reflecting aspects of text content, genre, authorship, etc.[16] Hearst (1999) argues that text categorisation is no text (data) mining, since it simply assigns texts to *predefined* categories without exploring "new information". The author shares this evaluation. But if the requirement is abandoned that these categories are predefined and if they are explored on their own, a text mining task can be seen to arise.
- **Text summarisation** has the aim of reducing textually coded information to its "essential points" based on intermediate, content based text representations.[17] The resulting summarisations serve as a starting point for generating summary texts.

In the following a fourth type of text mining task is described which is called *text linkage*. This is done by first specifying semiotic aspects of text mining, namely by characterising the complexity of its special object: *natural language texts*.

---

[12] See Hearst (1999). The difference of structural and referential relations points to alternative conceptions of sign meaning.

[13] See Fayyard et al. (1996).

[14] The subsequent enumeration is due to the variety of text mining approaches and their aims discussed in literature. A comprehensive overview of these different approaches is lacking as yet.

[15] See Cowie/Lehnert (1996), Grishman (1997), and Wilks/Catizone (1999) for overviews of information extraction.

[16] See Aas/Eikvil (1999) for an overview of methods for text categorisation.

[17] See Endres-Niggemeyer (1998).

SEMIOTIC ASPECTS OF TEXT MINING

Natural language texts differ fundamentally from the categorically, clear-cut defined data structures of *data mining*. Data mining procedures generally operate on relational (possibly normalised) data structures (*tables*) in which every attribute (*column*) has a unique meaning explicitly determined by its inter- and intrarelational dependencies. Arguing that texts are (in comparison to these units) *unstructured*, as numerous text mining approaches suggest, is misleading. Evidently texts do not represent sets of words or sentences, but have *semantic structure* beyond the sentence level. The organisation of texts – their *texture* – is based, among other things, on cohesion and coherence relations of text-forming resources which are used to connect text components in order to express *semantic* as well as *pragmatic continuity*.[18] Whereas cohesion refers to the characteristics of a text being a linguistic unit, coherence comprises situational, cognitive and social aspects of text processing. According to the system theoretical approach of Strohner/Rickheit (1990), cohesion and coherence are not strictly separated so that the unifying term *coherence* can be used to cover both aspects of textuality. As a rule, coherence relations, which indicate constraints for the interpretation of texts and their components (with respect to a given context of situation and culture), are *not* surface structurally explicit.[19] This aspect of uncertainty and context sensitivity is accompanied by the "structural plasticity" of coherence relations:[20] to varying degrees, coherence relations exist between units of different levels of linguistic resolution, not necessarily restricted to text components of a specific type (e.g. sentence or paragraph level), size or text distance.[21] This sort of structural plasticity, which formal data structures *lack*, *serves* for the constitution of texts as the most efficient "data structure" for the dissemination of information.[22] Beyond this intratextual dynamics[23], a second kind of variability, which relates to the mutual dependence of texts and their social contexts, has to be distinguished. This can be outlined following Systemic Functional Linguistics (SFL, see Halliday 1991) which views natural languages as dynamic, probabilistic systems persisting on the background of permanent interaction within their social environments, where texts form the primary unit of interaction. As natural languages are embedded into larger social semiotic systems, texts are produced/received depending on social contexts stratified into the context of situation and culture- connected via genres and registers with the language system. *Registers* manifest variety according to use depending on situational context. They are differentiated according to the dimensions of *field* (i.e. course of events), *tenor* (social roles of the participants), and *mode* (symbolic organisation of the situation), see Halliday 1978). *Genres* correspond to the staging of social processes as part of the context of culture and may be linguistically (though not deterministically) manifested by means of generic, schematic structures (see Martin 1992).[24] Halliday (1978) defines registers as manifestations of the correlation of semiotic context and semantic system. A register is a configuration of meaning potentials (and linguistic resources), that are *typically* connected with a certain situation type, that is a specific configuration of field, tenor, and mode. As patterns of semantic resources typically associated with one or more situation types in a given speech community, registers and genres serve to connect texts with their social contexts. In other words, the constitutive dependence of social and linguistic system is paralleled by the interdependence of context and text as instances of situation types and text types (as linguistic patterns for encoding realisations of genres and registers), respectively. Thus, texts always have two contexts: the (sub-)system of meaning potentials underlying their production/interpretation (mediated through one or more text types) as well as the (abstract) situation type(s), whose (real) instances form the social context for texts functioning as communication units. The semiotic

[18] See Halliday/Hasan (1976).

[19] In this sense, it is more appropriate to speak of natural language texts as "implicitly" structured entities than to misleadingly suggest that texts are "unstructured".

[20] This specific complexity of coherence is tied to its dependence on the dynamics of text production/reception.

[21] See Halliday (1994) for examples of these kinds of coherence relations.

[22] The term *efficiency* is used in the sense of Frege (1994), who states that natural languages are – in contrast to formal, artificial languages – adaptable/flexible in order to cope with new, unforeseen situations. As texts are the primary linguistic unit by means of which language is realised, they must realise this property, too.

[23] This is not to claim that coherence relations are restricted to texts. On the contrary, they embed texts into contexts.

[24] To separate between genre and register is not to negate the possibility that genre can be subsumed under mode as one dimension of context of situation, as Halliday (1978) suggests in contrast to Martin. This dispute does not affect the model proposed in the following.

complexity of this twofold contextual embedding relates to the fact that texts make not only use of meaning potentials and linguistic realisation patterns, but also confirm, modify or even constitute such regularities. Furthermore, because of the interdependence of social and language system text processing also affects what is conceived as social context.[25]

To summarise, coherence relations do not only connect components of the same text – constituting intratextual cotexts (i.e. textual contexts) for their interpretation – but also embed texts into social contexts by realising generic structures and registers. In other words, coherence relations do not only allow the same text component to be interpreted differently in different texts, but also the same text to function in different contexts. Finally, the co-realisation of the same or similar genres or registers is a source for intertextuality[26] so that coherence relations can be "exploited" for exploring intertextual relations.

As the scientific discipline for the study of signs, their constitution, and usage, *semiotics* analyses the complexity of signs in general.[27] On the background of the considerations outlined so far, the following aspects of complexity can be distinguished:

- **Toward the dynamics of sign meanings**: depending on different (cognitive, situative, or social) contexts, the same sign can have different meanings (*context sensitivity*). As a result of processes of stabilisation/perturbation of such regularities, signs, their properties and relations are enforced, inhibited, they emerge or disappear (*variability* with respect to language subsystems). Regarding the same text, the value of a quantitative text characteristic may vary in the course of text production/reception (*variability* with respect to single text systems). As Peirce (1983) points out, signs have a tripartite, relational organisation, according to which they enter into recursive processes of sign interpretation, called *semiosis*. The interpretation relations involved in such processes determine the signs' vague identities, their borders, properties and relations (*vagueness*).
- **Polyfunctionality and synergy**: the same (type or class of) sign(s) can serve for different (text forming) functions. Similarly, the same function can be served by different (types or classes of) signs (polyfunctionality). Signs cooperate and compete with respect to their (text forming) functions. The function of a (type or class of) sign(s) may enforce or suppress the function of other (types or classes of) signs (synergy).[28]
- **Realisation and covariation**: text signs *realise* genres and registers mediating between social and language system. As outlined above, SFL does not conceive this dependence as unidirectional instantiation, but as a sort of co-evolvement referred to as *redounding*. In other words, changes in context are correlated with changes in texture; there exists a systematic *covariation* of contextual and textual features (see Lemke 1995). As a consequence, context is seen as a semiotic entity in SFL, which does not simply serve for the disambiguation of textual components, but whose constitution depends on its part on sign usage.

From a semiotic point of view, processes of sign usage can modify those conditions, underlying their realisation. As a consequence, semiotics focuses on processes of sign (meaning) *constitution* instead of presupposing linguistic knowledge for the description of signs. According to the specific complexity of (text) signs, a significant difference of methods applied in text and data mining is expected. Furthermore, a methodically concretised computational semiotics appears to serve as a convenient framework for the specification (and, where appropriate, solution) of the task of *text (sign)* mining. *Computational semiotics* (CS) has been proposed as a scientific discipline in order to account for the intersection of *semiosis* and *computation*.[29] A computational semiotic model of sign processing is a *procedural model* reflecting the aspects of complexity outlined above. It comprises a representational format which

- separates representations of language system, social contexts, and texts,
- allows to model dynamics of context and language change so that its representations are modified, if the underlying data basis is changed, and
- enables to model the vagueness of semiotic phenomena by distinguishing gradual memberships of linguistic and contextual relations.

---

[25] See Halliday (1978).

[26] See Lemke (1995).

[27] See Peirce (1983).

[28] The synergetic aspect of self organization based on cooperating/competing linguistic processes which aim at adapting the properties of functional equivalents in order to optimally serve for cooperating/competing language needs is analysed in *synergetic linguistics*. See Köhler (1990).

[29] See Clarke/Mehler (2000).

Models in CS have to be formalised in algorithmic terms (procedures), which allow to delegate them to a computer. Results produced by these implementations are bound to the cotextual and contextual information underlying their processing. One does not claim to produce "representative" models regarding the totality of the (sub-)languages under consideration since any model of semiotic phenomena may change immediately if its context is changed. CS models are falsifiable regarding their theoretical background, procedural claims, and the structures they produce. With respect to the task of text mining, computational semiotics seems to be an adequate framework because it addresses the exploration of sign usage regularities in a computational model.[30] This insight can now serve as a semiotic reconstruction of text linkage as a task in text mining.

SYSTEMIC FUNCTIONAL TEXT LINKAGE

Text linkage refers to the exploration of implicit, content based, context sensitive relations of texts and their annotation as typed links in corpora organised as hypertexts. The aim of this section is to show that text linkage can be seen as a further task in text mining which naturally gives rise to a semiotic understanding as proposed in the last section.

Evidently, the task of automatically exploring intertextual relations and their representation as links in hypertext has not been solved yet. This lack correlates with the flexibility of human information processing: depending on a user's varying information needs as well as cognitive, situative, and social contexts the same text can be interpreted or at least evaluated differently with respect to its relevance. As a matter of principle, this context sensitivity of information processing implies that a text corpus does not have a predefined, deterministic hypertext structure:[31] different users with different information needs and contexts may find different intertextual relations to be relevant to be explicitly represented as links. Nevertheless, automatically creating text links is a crucial demand, because the amount of documents available online is rapidly increasing, so that an exploration problem (as described in the introductory section) arises.

Text linkage is an explorative task in the sense that it aims at answering the question which (type of) intertextual relation holds between which texts in which (type of) context. With respect to the variation of tasks (as a dynamic user perspective on the same text collection) text linkage aims at answering the question which texts are relevant for supporting which task (goal, or information need) subject to which context.

[30] Computational semiotics could be understood as a *theoretical* basis of text mining as a *technological* discipline.

[31] This has already been stated elsewhere. See Mehler (1999).

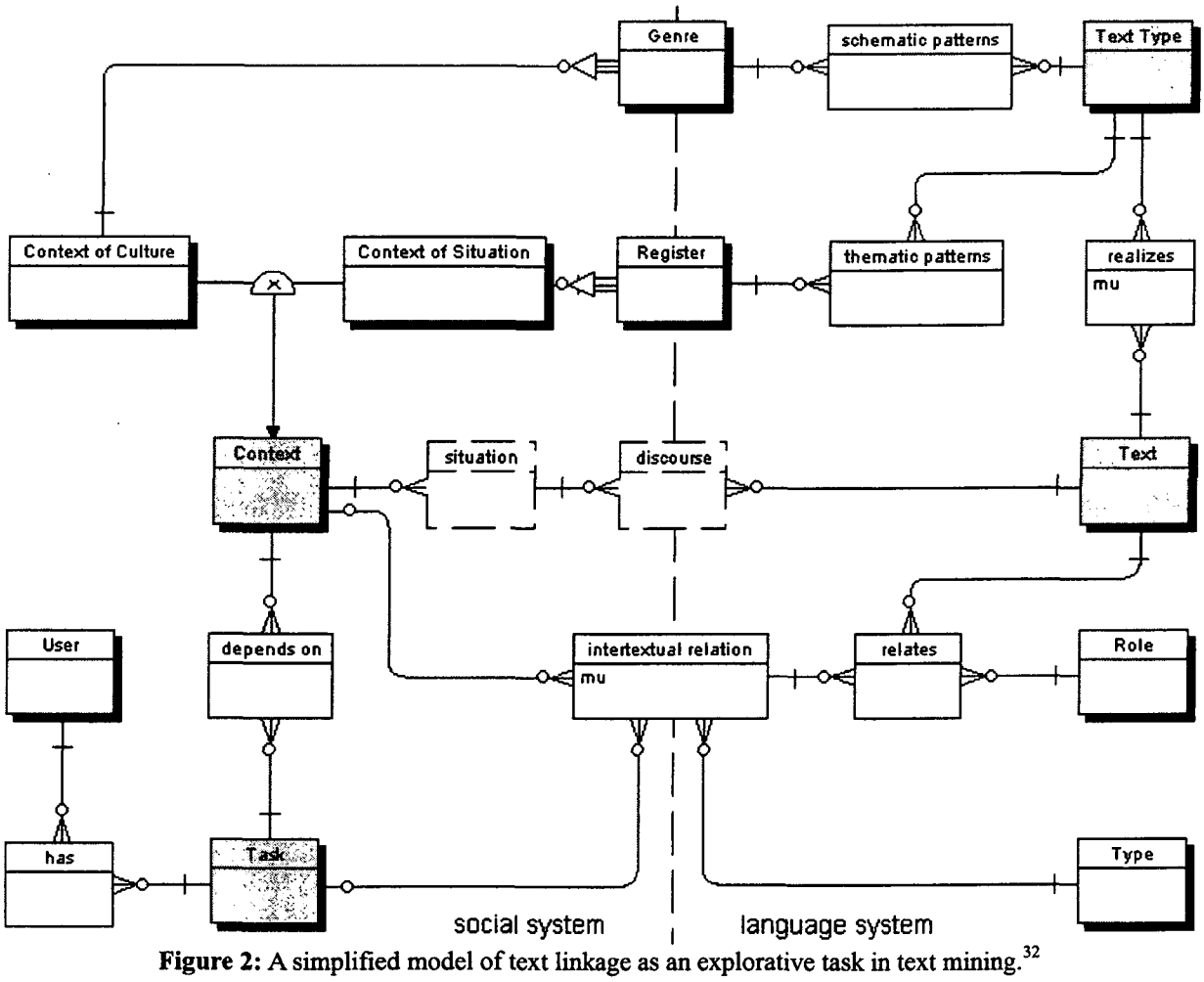

Figure 2: A simplified model of text linkage as an explorative task in text mining.[32]

In order to illustrate this, the data model in figure (2) comprises relevant categories affected by text linkage.[33] The basic building blocks are (social) CONTEXT with children CONTEXT OF CULTURE (C.O.C.) and CONTEXT OF SITUATION (C.O.S.), further: GENRE and REGISTER, TEXT, TEXT TYPE, and TASK:[34] GENRE and REGISTER, which are dependent on C.O.C. and C.O.S., respectively, are related with TEXT TYPE. An instance of TEXT TYPE is a system of linguistic patterns for realising one or more genres and/or registers (see SCHEMATIC PATTERNS and THEMATIC PATTERNS for modelling a many to many relation between TEXT TYPE, GENRE, and REGISTER).[35] The entity TEXT TYPE serves for the representation of thematic and generic heterogeneity of texts and its linguistic manifestation. In diagram (2) each text (see entity TEXT) is seen to belong (to varying degrees – see attribute MU representing membership values □) to one or more text types (see relation REALIZES). In the framework of text

[32] The conceptual model in figure (2) uses boxes and lines to represent entities and relationships, respectively. Lines are either straight (unique/one) or branching (multiple/many) in order to model the complexity of relationships. Furthermore, circles are used to represent optional relationships, whereas vertical bars are used to model mandatory relationships. Different styles of (shaded or dashed) boxes are used for highlighting objects. To exemplify this: the relationship between GENRE and SCHEMATIC PATTERNS means that a genre has $n \geq 0$ schematic patterns, whereas a schematic pattern belongs (is dependent on) exactly one genre (there is no schematic pattern without a corresponding genre). This example shows that the conceptual model in figure (2) simplifies entities and their relationships that need to be specified in a more detailed model.

[33] For the sake of simplicity, attributes have been skipped.

[34] SITUATION refers to instances of CONTEXT, i.e. real situations as instances of situation types. In the same situation one or more texts may be processed (as DISCOURSE).

[35] The entity THEMATIC PATTERNS models ideational, interpersonal, and textual meaning components, associated with field, tenor, and mode as dimensions of situation types. For the purpose of clarity, these and many other details are skipped in diagram (2).

linguistics[36], REALIZES can be seen as an extensional representation of text sorts, whereas TEXT TYPE serves as an intensional characterisation of texts. In contrast to genres and registers, text types are part of the language system. This is indicated by a vertical dashed line that separates the social system from the language system. Genres and registers serve to bridge both systems as indicated in diagram (2). Further, there is an entity TASK related to USER and CONTEXT (via DEPENDS ON). From the perspective of text mining, the most important entities are INTERTEXTUAL RELATION and RELATES, which serve to model any kind of homogeneous or heterogeneous intertextual relation classified by instances of TYPE. ROLE serves to determine the arguments of instances of INTERTEXTUAL RELATION (in the case of similarity relations, ROLE allows for example to model instances of asymmetric similarity). With respect to the entity TYPE, at least two types of intertextual relations can be distinguished:[37]

1. **Similarity**: *two texts may be intertextually related, if they are similar with respect to field, tenor, mode or generic structure (staging) they are realising.* Suppose, for example, two registers relating to similar fields (e.g. "natural disaster" and "environmental disaster"). Because of the similarity of their fields, both texts are expected to realise similar lexicogrammatical structures (e.g. similar lexical choices like "damage", "destruction", "number of the dead", etc.). This situation is illustrated in figure (3) by means of a commuting diagram, where $\square$ denotes similarity, ∇ stands for the realisation of register R_i by means of text x_i, and $\leftrightarrow$ denotes the intertextual relation of x_i, x_j.
2. **Dependence**: on the background of the same (macro) genre or register (realised by a system of intertextually linked texts) two texts may be linked because of their generic or thematic dependence. Suppose for example an organisation dealing with the management of working flows (social context) accompanied by technical documentation (generic structure). Suppose now a text describing a specific working flow in general and a second text dealing with any stage in this process. Obviously, processing the latter document depends on processing the former. This kind of intertextuality goes beyond similarity, since it is based on generic integrity of both texts: they are integrated into the same social process. The same holds for text linkage based on macro registers, where several texts are linked based on their contribution to the same field (tenor and mode) for realising the same instance of an homogeneous register.[38]

The differences between both types of intertextual relations can be exemplified as follows:

- Suppose two scientific texts in the area of cognitive linguistics which describe different experiments (comprising the following stages: *hypotheses*, *tests*, *results*, *prospect*, etc.) with *unrelated* contents. In this case, an intertextual relation holds between both texts based on their generic similarity (see figure 3.b). Suppose now two texts of the same area are given: whereas the first is supposed to describe an experiment, the second is a theoretical statement concerning the topics dealt with in the first text. In this case, a register based similarity is observed (because of related, though not identical fields) and hence a topic based intertextual relation is stated (see figure 3.a). Although both texts are possibly realised by totally different linguistic means (wording, etc.), they are seen to be semantically similar due to the realisation of related registers. Clearly, both kinds of similarity can be combined.
- Suppose now, a macro genre of series of experiments and two texts dealing with consecutive experiments of the same series is given. In this case, we observe a (generic) dependence relation between both texts.

As a text mining task, text linkage aims at answering the question which texts are connected by which type of relation, possibly dependent on which task and which context. The results of this exploration are represented by instances of INTERTEXTUAL RELATION (and its companions), which is optionally related with TASK and CONTEXT. Moreover, text linkage aims at exploring the relation of genre, register and texts. It asks which texts realise which registers/genres and whether there is a new register/genre to be introduced or already established registers/genres have to be merged. Because of the dependence of text links on higher level genre and register relations – and since these entities are not (strictly) presupposed –, text linkage can be seen as a special kind of "text categorisation" in which the system of "categories" is explored and dynamically maintained during processing of text corpora. Furthermore, genres and registers are not just unstructured categories, but have structure for the classification of textual macrostructures. In this sense, text linkage radically departs from text categorisation.

---

[36] See Heinemann (2000).

[37] This enumeration clearly does not claim to be complete, neither with respect to the range, nor to the details of possible text relations.

[38] If there is knowledge about the relation DEPENDS ON between TASK and CONTEXT, the dependence of different tasks – see the introductory section – can be mediately explored by means of the exploration of relations of registers (or genres).

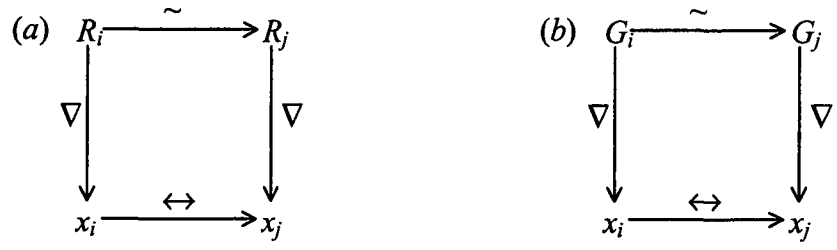

Figure 3: Intertextual dependence of two texts based on realising dependent registers (a) or genres (b).

As outlined so far, text linkage can be seen to serve as the explorative basis for automatically creating hypertexts from text corpora of heterogeneous genres and registers. Furthermore, being based on SFL, it allows the development of a linguistic typology of links. Bucher (1999) uses a generalised schema, in which a link *L* connects a node *a* with a node *b* dependent on *C*, thereby pointing to different tasks of link generation as outlined in table (1).

Task	Characteristic question	Constituent
Identification	**Which elements serve as a link source?**	*L*
Reference	**Which aspect of the source document is linked?**	*a*
Sequel	**Which aspect of the target document is linked?**	*b*
Typology	**Which type of link is given?**	*C*

Table 1: Dimensions of links in hypertexts according to Bucher (1999).

Of special interest is constituent *C*, which refers to the question of what type of link *L* is supposed to instantiate. Based on the understanding that links in hypertexts are realisations of intratextual and intertextual relations, the task of typology can be reformulated in terms of the (coherence providing) linguistic resources, the type of context involved, and the integrity or stability of links.[39] In other words, links in hypertext are seen to be stratified with respect to the linguistic organisation (cohesion) as well as situational (register based), and generic coherence of the texts to be linked. In this sense, links in hypertext can be seen as intertextual "coherent ties", whose strength depends on the threefold coherence of the texts linked. As links enter into hypertext paths, a further dimension of link typology has to be taken into account: purely associative links of text pairs need to be distinguished from chains of links, which constitute higher level hypertext structures.[40] Consequently, the typology for hypertext links distinguishes at least two dimensions:

1. The dimension of contextuality refers to the linguistic, situational and cultural basis of links.
2. The dimension of path sensitivity refers to the scope of links. It relates to the question of whether they are only used to link (associate) text pairs or whether they serve to build coherent paths or chains of interlinked texts.

Dimensions	COHESION	REGISTER	GENRE
ASSOCIATIVITY (of text pairs)	(1) associative link	(2) register associativity	(3) schematic associativity
Path sensitivity (of text chains)	(4) cohesive path	(5) situationally coherent hypertext path	(6) generically coherent hypertext path

Table 2: A two-dimensional systemic functional typology of text linkage.

[39] The link typology outlined in the following is not restricted to intertextual links, but comprises intratextual links as well.

[40] See Kuhlen (1991), who already discusses higher level link structures in hypertext for the representation of thematic and/or rhetoric progressions.

On this background, a provisional link typology can be outlined as given in table (2). At the lower end of this matrix purely associative links are distinguished – the typical case of links produced in the area of automatic hypertext construction – which neither take path, nor the context of register or genre into account. The construction of (lexically) cohesive as well as path sensitive links is described in Mehler (2000). The main idea of this approach is: instead of manifesting context-free associations, so called cohesion trees are used in order to model context priming effects based on the lexical organisation of consecutively linked texts. As a consequence, cohesion trees allow the perspective evaluation of numerically represented text similarities. Both kinds of links (i.e. link type (1) and (4) in table 2) are based on automatic corpus processing neglecting higher level genre and register dependencies. More coherent links are produced if these context layers are taken into account. The combination of genre or register with the concept of path sensitivity refers to links as constituents of hypertext paths connecting texts realising the same genre or register (see link types (5) and (6) in table 2). In the course of automatically generating hypertexts from newspaper articles, an example of generic coherent hypertext paths would be a sequel of dialogically ordered thesis and replies of a small set of participants discussing the same topic. An example would be the philosophical debate about gene technology in German feature pages initiated by the philosopher Sloterdijk). Obviously, associative, path insensitive links are of lowest stability as they vary according to the underlying text corpus: to enlarge this corpus may rapidly modify the strength of text links as the representations of meaning regularities of text components change due to the processing of new texts. In contrast to this, the combination of register and generic coherence extends a link's validity/stability. Thus, if hypertext construction aims at a higher level of readability or comprehensibility, the combination of cohesion, register, and genre stable links is needed. But since registers and genres are "shaped" by means of texts, the enlargement of the corpus base may lead to a modification of such links as well in the longer run.

These considerations induce a four-layer architecture for the construction of so called *systemic functional hypertexts*: First, the reconstruction of genre structures occurs. On this level, the constituency, staging as well as association and dissociation of genres is modelled. Second, on the level of registers, networks of registers as well as accessibility constraints of situation types are described. Third, the language layer models classes of texture forming resources, their paradigmatic organisation, and linguistic realisation patterns. Fourth, the text layer describes context sensitive intra- and intertextual relations represented as hypertext links. Hypertexts of this kind are stratified hypertexts: the first two layers consist of interlinked nodes, which dominate sets of texts as realisations of the corresponding genre or register representation. As explained above, this dominance relation does not induce an equivalence relation over the corpus of texts.

REQUIREMENTS ANALYSIS

As a theoretical background, computational semiotics points to specific requirements for solving the problem of text linkage as a task in text mining. In this context, the following requirements for any implementation of a text linkage procedure can be distinguished:

- **Text meaning and coherence**: as a starting point for text linkage, representations of aspects of text meaning are needed, which reflect coherence relations as the fundamental organisation unit of natural language texts. These representations have to make explicit to which degree they reflect certain types of textual coherence and certain aspects of text meaning.
- **Implicit text relations**: the mining procedure should be able to link texts, even if they share only few or no components at all, but deal with related or similar topics or realise related genres or registers, respectively. The significance of this requirement can be exemplified by looking at texts in different languages, which are totally different in terms of surface structure, but may be similar regarding their content.
- **Reconstructive, explorative corpus analysis**: the decision to link texts is based on knowledge reconstructed or even explored from the corpora to be converted into hypertext. Any text linkage has to take the specifics of signs and their usage in the corpora into consideration. This does not mean that, initially, text linkage is knowledge free since it possesses at least procedural knowledge of how to acquire relevant information. Furthermore, any factual knowledge that the system has already disposed of, should be the result of reconstructive, explorative corpus analysis of its own.[41]
- **Representational format and context model**: the specific complexity of texts requires a representational format, which allows to model the vagueness of text relations with any resolution, disposes of context correlates and is adaptable to the analysis of texts of previously unknown registers and genres. In addition, the format has to be sensitive to new cotextual and contextual evidence, which may justify modifications of the representations already produced.

[41] This requirement is certainly a difficult, if not an unrealistic one at present. Nevertheless, there exist computational models, which aim at reconstructing morpho-syntactic knowledge from large text corpora. See for example Medina Urrea (2000).

The fulfilment of these requirements enables *transparent* knowledge acquisition, since factual (extensional) knowledge, whose acquisition is tied to human expertise, is replaced by intensional, procedural specifications.

PROSPECTS

The considerations outlined so far are rather programmatic: they have tried to clarify the task of text mining from the perspective of computational semiotics with special focus on the task of text linkage as an explorative basis for the automatic construction of hypertexts. It was the aim of the paper to show that text mining cannot rely on classical methods of information retrieval or knowledge engineering since its object, i.e. natural language texts, have a complexity which goes beyond the complexity of classical data structures. Clearly, this aim could only be tentatively, motivationally realised leaving the proof of all necessary evidence and examples to forthcoming papers. Thus, future work will aim at giving procedural definitions of genres, registers, text types, and text forming resources. Special emphasis will be on the corpus analytic reconstruction of these entities. Moreover, a formal model is needed for the complete formalisation of systemic functional hypertexts, thereby making formally explicit such (more or less vague) terms as context sensitivity, variability, vagueness, polyfunctionality, realisation and covariation. Finally, the model has to prove its computability and applicability.

REFERENCES

Aas, K.; Eikvil, L. (1999): Text Categorisation: A Survey. Norwegian Computer Center, **Report No. 941, 1-37.**

Bucher, H.-J. (1999): Die Zeitung als Hypertext. Verstehensprobleme und Gestaltungsprinzipien für Online-Zeitungen. In: Lobin, H. [ed.]: **Text im digitalen Medium: linguistische Aspekte von Textdesign, Texttechnologie und Hypertext-Engineering.** Opladen: Westdeutscher Verlag, 9-32.

Clarke, R.; Mehler, A. (1999): Theorising Print Media in Contexts: A Systemic Semiotic Contribution to Computational Semiotics. In: **Proceedings of the 7th International Congress of the IASS-AIS: International Association for Semiotic Studies - Sign Processes in Complex Systems**. University of Technology, Dresden, Germany, October 6-11, 1999.

Cowie, J.; Lehnert, W. (1996): Information Extraction. In: **Communications of the ACM, 39(1),** 80-91.

Endres-Niggemeyer, B. (1998): **Summarizing Information.** Berlin [a.o.]: Springer.

Fayyard, U.; Piatetsky-Shapiro, G; Smyth, P. (1996): The KDD Process for Extracting Useful Knowledge from Data. In: **Communications of the ACM, 39(11),** 27-34.

Frege, G. (1994): **Funktion, Begriff, Bedeutung. Fünf logische Studien.** 7. edition; Göttingen: Vandenhoeck & Ruprecht.

Grishman, R. (1997): Information Extraction: Techniques and Challenges. In: Pazienza, M. T. [ed.]: **Information extraction: a multidisciplinary approach to an emerging information technology; international summer school, SCIE-97**, Frascati, Italy, July 14-18, 1997. Berlin [a.o.]: Springer, 10-27.

Hahn, U.; Schnattinger, K. (1998): Towards Text Knowledge Engineering. In: **Proceedings of the 15th National Conference on Artificial Intelligence & 10th Conference on Innovative Applications of Artificial Intelligence**, Madison, Wisconsin, July 26-30, 1998. Cambridge: AAAI Press, MIT Press, 524-531.

Halliday, M. A. K. (1978): **Language as Social Semiotic**. London: Edward Arnold.

Halliday, M. A. K. (1991): Towards Probabilistic Interpretations. In: Ventola, E. [ed.]: **Functional and Systemic Linguistics.** Berlin [a.o.]: de Gruyter, 39-61.

Halliday, M. A. K. (1994): **Introduction to functional grammar**. London: Edward Arnold.

Halliday, M. A. K.; Hasan, R. (1976): **Cohesion in English**. London: Longman.

Hearst, M. A. (1999): Untangling Text Data Mining. In: **Proceedings of ACL '99: the 37th Annual Meeting of the Association for Computational Linguistics**, University of Maryland.

Heinemann, W. (2000): Textsorte – Textmuster – Texttyp. In: Brinker, K.; Antos, G.; Heinemann, W.; Sager, S. F. [eds.]: **Text- und Gesprächslinguistik. Linguistics of Text and Conversation**. Berlin [a.o.]: de Gruyter, 507-523.

Klir, G. J.; Folger, T. A. (1988): **Fuzzy Sets, Uncertainty, and Information**. Englewood: Prentice Hall.

Köhler, R. (1987): Systems Theoretical Linguistics. In: **Theoretical Linguistics 14(2/3),** 241-257.

Kuhlen, R. (1991): **Hypertext: ein nichtlineares Medium zwischen Buch und Wissensbank**. Berlin [u.a.]: Springer.

Lemke, J. L. (1995): Intertextuality and Text Semantics. In: Fries, P. H.; Gregory, M. [eds.]: **Discourse in Society: Systemic Functional Perspectives. Meaning and Choice in Language: Studies for Michael Halliday**. Norwood: Ablex Publishing, 85-114.

Lobin, H. (2000): **Informationsmodellierung in XML und SGML**. Berlin [a.o.]: Springer.

Martin, J. R. (1992): **English Text. System and Structure**. Philadelphia [u.a.]: John Benjamins.

Medina Urrea, A. (2000): Automatic Discovery of Affixes by Means of a Corpus: A Catalog of Spanish Affixes.

In: **Journal of Quantitative Linguistics 7(2),** 97-114.

Mehler, A. (1998): Toward Computational Aspects of Text Semiotics. In: **Proceedings of the 1998 IEEE ISIC/CIRA/ISAS Joint Conference on the Science and Technology of Intelligent Systems**. Piscataway: IEEE/Omnipress, 807-813.

Mehler, A. (1999): Aspects of Text Semantics in Hypertext. In: Tochterman, K.; Westbomke, J, Wiil, U. K., Leggett, J. J.: **Hypertext '99. Returning to our diverse roots. Proceedings of the 10th ACM Conference on Hypertext and Hypermedia**. Darmstadt: ACM, 25-26.

Mehler, A. (2000): Text Mining with the Help of Cohesion Trees. Appears in: **Proceedings volume of the annual conference of GfKl** at the University of Passau, Germany, 2000.

Peirce, C. S. (1983): **Phänomen und Logik der Zeichen**. Frankfurt am Main: Suhrkamp.

Rieger, B. (1989): **Unscharfe Semantik: die empirische Analyse, quantitative Beschreibung, formale Repräsentation und prozedurale Modellierung vager Wortbedeutungen in Texten**. Frankfurt am Main: Peter Lang.

Sparck Jones, K. (1999): What is the Role of NLP in Text Retrieval? In: Strzalkowski, T. [Hrsg.]: **Natural Language Information Retrieval**. Dordrecht [u.a.]: Kluwer, 1-24.

Strohner, H.; Rickheit, G. (1990): Kognitive, kommunikative und sprachliche Zusammenhänge: Eine systemtheoretische Konzeption linguistischer Kohärenz. In: **Linguistische Berichte** *125*, 3-24.

Wilks, Y.; Catizone, R. (1999): Can we Make Information Extraction more Adaptive? In: Pazienza, M. T. [ed.]: **Information Extraction. Towards Scalable, Adaptable Systems**. Berlin [a.o.]: Springer, 1-16.

Zadeh, L. A. (1997): Toward a Theory of Fuzzy Information Granulation and its Centrality in Human Reasoning and Fuzzy Logic. In: **Fuzzy Sets and Systems 90**, 111-127.