*Journal of Universal Computer Science, vol. 8, no. 10 (2002), 924-943*
*submitted: 24/4/02, accepted: 15/10/02, appeared: 28/10/02* 

# Components of a Model of Context-Sensitive Hypertexts

Alexander Mehler
(University of Trier, Germany
mehler@uni-trier.de)

**Abstract:** On the background of rising Intranet applications the automatic generation of adaptable, context-sensitive hypertexts becomes more and more important [El-Beltagy et al., 2001]. This observation contradicts the literature on hypertext authoring, where Information Retrieval techniques prevail, which disregard any linguistic and context-theoretical underpinning. As a consequence, resulting hypertexts do not manifest those schematic structures, which are constitutive for the emergence of text types and the context-mediated understanding of their instances, i.e. natural language texts. This paper utilizes *Systemic Functional Linguistics* (SFL) and its context model as a theoretical basis of hypertext authoring. So called *Systemic Functional Hypertexts* (SFHT) are proposed, which refer to a stratified context layer as the proper source of text linkage. The purpose of this paper is twofold: First, hypertexts are reconstructed from a linguistic point of view as a kind of supersign, whose constituents are natural language texts and whose structuring is due to intra- and intertextual coherence relations and their context-sensitive interpretation. Second, the paper prepares a formal notion of SFHTs as a first step towards operationalization of fundamental text linguistic concepts. On this background, SFHTs serve to overcome the theoretical poverty of many approaches to link generation.

**Keywords:** Hypertext Authoring, Context Modelling, Coherence Relations, Systemic Functional Linguistics
**Categories:** H.3.1, H.3.3, H.5.4, I.2.7, I.7

## 1 Introduction

The majority of approaches to automatic authoring of hypertexts concentrate on binary links where the central unit to be optimized is the *similarity of pairs of texts* normally computed on the basis of the vector space model [Allan, 1997, Chen, 1997, Salton et al., 1994, Wilkinson and Smeaton, 1999]. Although higher level link concepts (e.g. link typing, paths, composite nodes) have already been introduced and successfully applied in early stages of hypertext [Halasz, 1988, Kuhlen, 1991, Zellweger, 1989], the literature still focuses on the narrow context of binary links [Agosti et al., 1997, Agosti and Smeaton, 1996, Allan, 1997, Salton et al., 1994]. As a consequence, the majority of existing systems misses any theoretical-linguistic grounding of text linkage. A more recent exception are topic tracking systems [Carthy and Smeaton, 2000], which explore temporal order of text production [Dalamagas and Dunlop, 1997] as well as lexical chaining [Green, 1998], i.e. the linkage of systematically related words participating in sense relations (e.g. synonymy, antonymy, hyperonymy). Nevertheless, lexical chains are used to optimize binary text links [Ferret, 2002], whereby dependencies of indirectly linked nodes manifesting paths as a kind of *hypertextual* context are disregarded. Moreover, although topic tracking systems aim at ordering texts dealing with the same event, *context layers* which control text linkage top-down are missed.

This paper departs from this approach. It radically shifts the perspective from context-insensitive, binary links to context-sensitive hypertextual structure formation, whereby the context model of *Systemic Functional Linguistics* (SFL) [Halliday and Hasan, 1976, Halliday, 1994, Martin, 1992, Ventola, 1987] is utilized as a text linguistic basis of text linkage. As a consequence, links are evaluated with respect to coherence relations they manifest, which on their turn are interpreted with respect to dependencies of context units they textually realize. This approach follows the evident fact that text understanding is context-sensitive: Depending on readers' varying cognitive, situational, and social contexts the same text may be interpreted differently. Text linkage does not suspended this dependency, but intra- and intertextual relations have (and thus also their hypertextual manifestations) a context-sensitive semantics. Applied to hypertext authoring, this means that context restricts text linkage. Corpora do not have pre-established, deterministic hypertext representations. Rather, these vary with respect to the operative context so that context-sensitive hypertext authoring becomes indispensable.

The remainder of the paper is organized as follows: [Section 2] briefly describes genres and registers as main constituents of SFL's context model. The essential building blocks of SFHTs are described in [Section 3], substantiated in [Section 3.1-3.2] and finally exemplified in [Section 4]. [Section 5] gives some conclusions.

## 2 Text and Context

Leaving out the details of language modelling in the framework of SFL this paper concentrates on two central implications of SFL for text linkage: (i) Links are manifestations of intra- and intertextual coherence relations[1] with a twofold contextual support, on generic and registerial level. (ii) Being conditional on this contextual embedding, links vary with respect to their contribution to the constitution of hyper-text structures (e.g. paths, composite nodes, networks of textual units). This point of view is justified by the indispensable contextual embedding of natural language texts which is observable by means of uniformities occurring across different situations of text production/understanding. According to these regularities, *types of contexts* can be distinguished, which differ with respect to

[1] We regard cohesion relations as special cases of coherence relations, whose arguments are restricted to be textual units.

their influence on text production/understanding, whereby the most general dimension of context classification regards *textual mode*, or more specifically: the distinction between speech and writing [Biber, 1991].

Regarding situative context, *Situation Semantics* [Barwise and Perry, 1983] reflects this regularity by describing actual situations as instantiations of abstract situation types. SFL instead concentrates on the socio-linguistic aspect of this type-instance relation (leaving out any formalization of its context model), whereby two fundamental dimensions of contextual structuring are distinguished: *genre* and *register* [Ventola, 1987]. Genre, or the context of social action, is referred to by a model of the staging of social processes and their patterns of linguistic realization. Genres manifest variety according to this staging and may be linguistically manifested by *schematic structures*. As an example consider the genre of giving a talk, where dependency relations (e.g. a hypotheses formation stage is followed by a stage of evidence giving), constituency relations (e.g. an introduction stage may comprise a problem setting and a hypotheses formation stage) and typological relations (giving a talk is a kind of speech genre), can be distinguished.[2] A central claim of SFL is now that the staging of genres correlates with significant changes in the choice of linguistic units. Registers, on the other hand, manifest situational variety and are seen to be structured by three parameters (so called *register variables*): *field* (referring to what is going on; what is described), *tenor* (referring to how it is evaluated by whom in which role), and *mode* (referring to how it is medially organized). Analogously to genres, SFL hypothesizes significant changes of linguistic realization patterns according to register change (the field of sports "primes" other lexical choices than the field of "classical music").

The structuring of context according to genres and registers and their role in text production/understanding can be explained by constructing a metaphor, where genres are generic workflow types abstracting from the concrete task to be solved, whereby registers specify these tasks, persons and their roles involved, etc. Clearly, genres and registers are only ideally orthogonal. Rather, the choice of a genre predetermines aspects of field, tenor, and mode. In any case, generic staging is rather comparable with the staging of workflows, whereby registers specify the instantiation of the stages involved.

On this background, texts always have two contexts: a subsystem of meaning potentials (i.e. a genre/register) underlying their production/interpretation and an instance of a context type in which these texts function as communication units. The complexity of this contextual embedding relates to the fact that

[2] Though this example suggests the existence of computational procedures for the identification of genres and their structuring, a set of criteria for the inter-subjective classification of texts according to such context units is yet not known. This is due to the general lack of operationalization (formalization and quantification) of text linguistic terms. As a consequence, a central task of implementing SFHTs will be the operationalization of these terms.

texts do not only instantiate genres and registers, but also confirm, modify or even constitute their organization. This relation of *mutual evolvement* of text and context is called *redounding* in SFL. It is a source of the non-deterministic predictability of text by context, and vice versa. From this perspective, context cannot longer be ignored, but becomes the proper basis of text linkage.

## 3 Systemic Functional Hypertexts

The definitional basis of Systemic Functional Hypertexts (SFHT) is given by a system of text linguistic terms and their appropriate computational linguistic operationalization (i.e. formalization and quantification). SFHTs are proposed as a computational linguistic format for the representation of contextually grounded intra- and intertextual relations as links in hypertext. They primarily focus on the *exploration* of links from text corpora and relate only secondarily on the systemic functional interpretation of already given, otherwise pre-established hypertexts. SFHTs are used as the underlying representational format of automatic hypertext authoring: they serve to specify the most general building blocks and their organization from a computational linguistic point of view. With respect to their operationalization two fundamental steps have to be distinguished:

- First, the system of theoretical linguistic terms seen to be constitutive for SFHTs has to be formalized in a way which identifies for each of these terms a representational correlate as a building block of SFHTs.
- Second, these basic terms, their representational correlates and relationships have to be quantified in a way, which allows their controlled, reproducible (semi-)automatic reconstruction or even exploration from natural language text corpora and possibly from other linguistic resources (e.g. lexica).

This paper contributes to the first of these two tasks by distinguishing indispensable constituents of SFHTs and their relationships on a very abstract level. Although this is done in formal mathematical terms, the formalization proposed is rather provisional in the sense that it requires several refinements and specifications of many of its yet provisionally introduced constituents in order to be usable as a starting point for quantification.

The text linguistic underpinning of SFHTs, which can only be hinted at, relates to an integration of the concept of coherence relation and the twofold contextual stratification as described in SFL. In order to sketch this integration, we argument as follows: Obviously, natural language texts are not sequences of unrelated sentences, but text constituents must cohere in order to form texts. Following the approach of [Knott and Sanders, 1998], we refer to the concept of *coherence relation* as a source of textual coherence: Processing a text, the reader builds a cognitive representation of its (e.g. propositional) content which aims to integrate the content representations of its parts, whereby coherence relations—possibly surface-structurally signalled by means of linguistic cues—specify how the constituent representations have to be integrated. A central implication from the point of view of context models as elaborated in text linguistics, as for example in terms of super-structures [van Dijk and Kintsch, 1983], is that surface-structurally signalled coherence relations help to identify the type of the text to be processed and thus to activate schematic knowledge restricting the identification and interpretation of succeeding spans and their coherence relations. In this sense, the distribution of coherence relations (and their resulting quantitative characteristics) are context-sensitive: the average distance of anaphoric references, for example, and the length of reference chains homophoric anaphora produce are expected to vary with the underlying text type (genre). It is this co-variation which is seen to underly the non-deterministic mutual predictability of context and texture as described in SFL. Following once more SFL, this paper describes context in terms of genre and register.

In other words: Dependent on the contexts in which they are produced, texts manifest at least in part genre- and register-specific structures of coherence relations (lexical and reference chains [Halliday and Hasan, 1989], (poly-)hierarchies of rhetoric relations [Mann and Thompson, 1988], etc.), whose semantics is restricted by the types of coherence relations they instantiate.[3] From a representational point of view, a coherence type specifies in general terms how the interpretation of a text span, connected by an instance of that type with another span of the same text, is restricted by the interpretation of the latter span (and vice versa as in case of paratactic relations).[4] The texture of a text as resulting from the coherence relations of its components and units of the context in which it is produced/understood does not only restrict the (con-)text specific interpretation of its components, but serves as a complex restriction for the integration of these interpretations into the content representation of the text as a whole.[5] But since text production/understanding is mediated by genres and registers, the interpretation of texts along intra- and intertextual coherence relations is conditional on these context types, too: whereas coherence types generally classify the semantics of coherence relations, it is their concrete interpretation which is dependent on instances of context types. Furthermore, different types of coherence relations are correlated with different lexicogrammatical choices as their preferred linguistic realizations, whereas genres and registers are correlated with

[3] In order to keep the formalism simple, we do not distinguish between text and discourse and thus define coherence relations, following [Mann and Thompson, 1988], as elements of cartesian products of the set of text spans. For a more detailed approach see [Mehler, 2002c]

[4] Anaphoric reference, for example, demands anaphora to have the same referential meaning as their antecedents.

[5] Take for example the the specific interpretation of the term *integration* in the present paper in contrast to its socio-cultural interpretation in other texts.

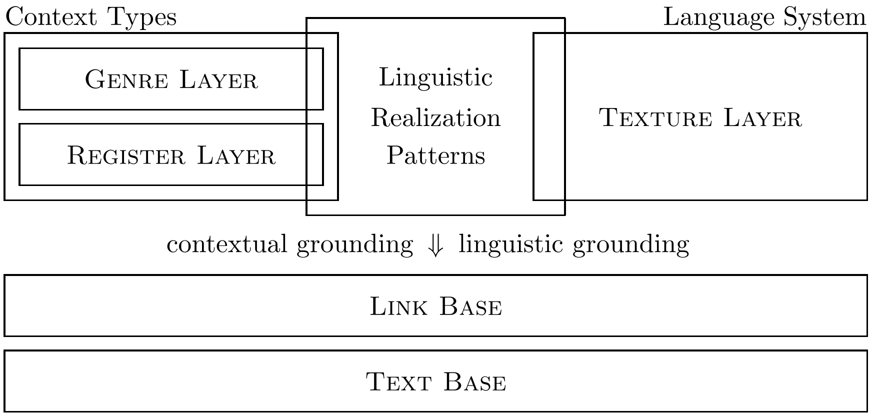


**Figure 1:** The layered architecture of SFHTs.

complex structures of these types of coherence relations, and thus correlate indirectly with patterns of linguistic units.[6]

From the perspective of hypertext authoring, two aspects of contextual embedding of coherence relations are important: Genres and registers do not only impose restrictions on the interpretation of single coherence relations[7], but the integration of these relations into complex discourse structures (e.g. thematic progressions) may be restricted by genre staging and register networking, too. In this sense, linguistically manifested context types are seen to be exploitable as a source of text linkage in hypertext, whereby SFHTs radically shift the perspective from barely linguistically interpreted links onto the level of linguistically and contextually grounded links and their integration into complex hypertextual structures (i.e. paths and composite nodes).

The (semi-)formal specification of SFHTs follows the line of their main building blocks: genres, registers, and texture forming resources. Admittedly, formal definitions reduce readability, but are indispensable when trying to clear the vagueness of text linguistic terms.

**Definition 1** A *Systemic Functional Hypertext* (SFHT) is an $n$-level hypertext $\langle GL, RL, TL, TB, LB, \ldots \rangle$, $n \geq 5$, which includes at least four main layers:

1. The *genre layer GL* models typology, constituency, and dependency relations

[6] A scientific discourse, for example, is more likely to realize complex argumentations realizing nested elaboration relations than informal speech and thus is more likely to contain words signalling elaborations.

[7] A coherence relation may connect, for example, two text spans realizing two succeeding stages or the lexical cohesion of both spans may serve as an indication of the similarity of the registers they instantiate.

of genres and their constituents (stages). It also includes representations of patterns of linguistic realization of generic units.

2. The *register layer RL* models typology, constituency, and dependency relations of registers and their constituents (instances of *field*, *tenor* and *mode*). It also includes representations of patterns of linguistic realizations of registerial units.

3. The *texture layer TL* models coherence relations as types of texture forming resources, their syntagmatic/paradigmatic relations and patterns of lexico-grammatical resources, whose instances realize the types in question.

4. The *text layer* models intra- and intertextual coherence relations as links, whose markers are identified with the text spans they connect. The text layer is implemented as a two-level system, where the *text base TB* organizes texts and their segments, whereas the *link base LB* organizes links and complex link structures. □

Whereas the first three layers describe different contextual and linguistic resources of text linkage, the last layer deals with the organization of concrete links. In order to simplify our formalism, linguistic realization patterns of genres and registers are modelled by a separate mapping as shown in [Fig. 1]. Now, the building blocks of SFHTs can be specified in more detail:

**Definition 2** Let $G$ and $S$ be finite sets of genre and stage labels, respectively. The set of *generic elements* is $X = G \cup S$. □

**Definition 3** The *genre layer* is modelled as a quadruple $GL = \langle X, \mathcal{D}, \mathcal{C}, \mathcal{T} \rangle$, where $X = G \cup S$ is a set of generic elements and $\mathcal{D}$, $\mathcal{C}$, $\mathcal{T}$ model dependency, constituency, and typological relations of generic elements, respectively:

1. $\mathcal{D} = (\langle S, \mathcal{E} \rangle_g \,|\, g \in G)$ is a family of fuzzy directed hypergraphs indexed by genres $g$ they define, where $\mathcal{E} \in \mathbb{F}(\{\langle \mathcal{A}, \mathcal{B} \rangle_d \,|\, d \in \mathbb{D}, \mathcal{A}, \mathcal{B} \in \mathbb{F}(S), \mathrm{hgt}(\mathcal{A}), \mathrm{hgt}(\mathcal{B}) > 0\})$ is a fuzzy set of fuzzy connected directed edges. $\mathbb{D} = \{\wedge\backslash\wedge, \wedge\backslash\vee, \vee\backslash\wedge, \vee\backslash\vee\}$ classifies edges $\langle \mathcal{A}, \mathcal{B} \rangle_d$ regarding their conjunctively/disjunctively joined input units $\mathcal{A} \in \mathbb{F}(S)$ and output units $\mathcal{B} \in \mathbb{F}(S)$, respectively. $\langle \mathcal{A}, \mathcal{B} \rangle_d \in \mathcal{E}$ is called *generic link* of *degree* $\mu_{\mathcal{E}}(\langle \mathcal{A}, \mathcal{B} \rangle_d)$ with respect to $g$. $\mathbb{F}(S)$ is the set of fuzzy sets over $S$.

2. $\mathcal{C} = (\langle S, E, g \rangle_k \,|\, g \in G, k \in \mathfrak{K})$ is a family of forests modelling concurrent, functionally divergent constituency structures of the same genre, where each tree of each forest is mapped by a function $f_{\mathcal{C}}$ onto a set of types of constituency structures $\mathbb{C}$ (e.g. *rhetorical structure*, *similarity based clustering*, etc.).

3. $\mathcal{T} = \langle X \cup \mathbb{T}, E' \rangle$ is a forest modelling typological relations of generic elements, where each leaf of each tree in $\mathcal{T}$ belongs to $X$ and all dominating nodes to the set of types of generic elements $\mathbb{T}$. □

[Def. 3] omits many details regarding the complexity of generic staging. It only serves as a *working definition*, which helps to formally narrow down main building blocks of SFHTs from the perspective of genres. This requires several comments:

1. According to [Def. 1], the genre layer includes representations of patterns of linguistic realization of generic units yet omitted in [Def. 3]. These patterns have to be concretized according to a specification of the kinds of linguistic structures represented as part of the texture layer.
2. Genres, whose constituents are dependent stages normally realized by text segments, have to be distinguished from *macro genres*, whose constituents are—as being autonomous genres on their own—realized by whole texts. Take for example the macro-genre of *online sports event presentation* as comprising the *live ticker*, *game background*, *game report*, and *short biography* genre. This distinction is omitted in [Def. 1], too.
3. The reference to the concept of fuzzy set excludes by no means probabilistic models of genre. On the contrary, since probabilistic models of syntactic structures prove to be effective in computational linguistics, comparable models of generic staging are preferred. In this sense, the reference to fuzzy sets has to be understood as nothing more than a generalization regarding different types of informational uncertainty.
4. Because of leaving out many constraints of generic structuring, [Def. 3] overgenerates generic structures.

Comparable remarks analogously apply to the definition of registers. Leaving out their complexity according to the variation of field, tenor, and mode, this paper concentrates on field-based register structuring and networking. This is done (somehow in analogy to the framework of ontologies) by supposing a set of elementary topic categories (as constituents of field—e.g. *economics*, *finance*, *sport*, *football*, etc.), whose dependency and constituency relations are referred to as main building blocks of registerial units (the automatic reconstruction of these categories will be decisive for implementing SFHTs). Without explicitly distinguishing these relations, hypergraphs are once more used as a representational format, where vertices represent—analogously to topics in Topic Maps [Widhalm and Mück, 2001]—elementary constituents, whereas typed, possibly directed edges model—comparable to sense relations as distinguished in WordNet [Fellbaum, 1998]—dependencies of these constituents. Furthermore, in order to model macro-registers as being built out of more elementary registers (e.g. the field of sports comprises the field of football, tennis, etc.) and starting from the concept of hyper-structure as introduced in [Baas, 1994], we refer to a recursive hypergraph definition which allows to model nested as well as overlapping registers. The generality of this definition guarantees independence with respect to existing approaches to ontology modelling as demanded by the rather theoretical investigations of this paper.

**Definition 4** The *register layer* is recursively defined:

1. $RL_0 = \langle V_0 \cup T, \mathcal{E}_0 \rangle$ is a fuzzy directed hypergraph, where $V_0$ is a set of topic labels, $T$ is a set of relation types and $\mathcal{E}_0$ is a fuzzy set over $\mathbb{F}(V_0)^2 \times T$ so that for each $e = (\mathcal{A}, \mathcal{B}, t), \mathcal{A}, \mathcal{B} \in \mathbb{F}(V_0), t \in T$, $\mu_{\mathcal{E}_0}(e) > 0$, $\mathcal{A}$ specifies the input and $\mathcal{B}$ the output nodes of the directed edge $e$. Undirected edges are derived from $\mathcal{E}_0$ and collected by $\mathcal{U}_0$ as follows: $\mathcal{U}_0 = \{\mathcal{A} \in \mathbb{F}(V_0) \mid \exists t \in T : \mu_{\mathcal{E}_0}((\mathcal{A}, \emptyset_{V_0}, t)) = 1 \vee \mu_{\mathcal{E}_0}((\emptyset_{V_0}, \mathcal{A}, t)) = 1\}$, where $\mathrm{hgt}(\emptyset_{V_0}) = 0$.
2. $RL_n = \langle V_n, \mathcal{E}_n \rangle$ is a fuzzy directed hypergraph with vertex set $V_n = V_{n-1} \cup \mathcal{E}_{n-1}$, whose edge set $\mathcal{E}_n$ is a fuzzy set over $\mathbb{F}(V_n \setminus T)^2 \times T$, whereby undirected edges are collected by $\mathcal{U}_n$ analogously to $\mathcal{U}_0$ in $RL_0$.

Now, the *register layer* can be defined as a hypergraph $RL = \langle V_n, \bigcup_{i=1..n} \mathcal{E}_i, R \rangle$ for some $n \geq 0$, where for each register $r \in R$ there exists an $i = \{0, \ldots, n\}$ so that $r$ is a connected sub-hypergraph of $RL_i$. □

Some remarks may help to rank this definition:

1. The recursive definition of $RL_i$, in which the edge sets of preceding layers become potential vertices of that hypergraph, is the primary instrument for modelling macro-registers and their nested, overlapping structures—somehow in analogy to scopes in Topic Maps. Moreover, this allows to model edges connecting different sub-hypergraphs and thus accessibility constraints between registers (the register of broadcasting rights, for example, makes accessible the registers of sports and mass media without being simply the union of these two registers).
2. Since the edge sets of the layers $RL_i$ allow to derive undirected edges, more general relations, as for example similarity clusterings of registerial units, can be modelled, too.

**Definition 5** The *texture layer* is modelled as a quadruple $TL = \langle K, O, S, P \rangle$, where $O$ models the hierarchical organization of coherence forming resources (i.e. coherence types), whereas $S$ and $P$ model syntagmatic and paradigmatic relations of linguistic realizations of coherence types, respectively:[8]

[8] For the sake of simplicity, the models of syntagmatic and paradigmatic relations of coherence types are omitted here.

1. $K = \{k_i \mid i \in \mathfrak{I}\}$ is a set of *coherence types* organized (comparable to systemic networks in SFL) as a connected directed acyclic graph $O = \langle K \cup \{\top\}, D\rangle$ with root $\top$ so that for each node $k_i \in K$ there exists a path connecting $\top$ with $k_i$.[9] Signatures of coherence types $k_i$ are symbolized as $\sigma(k_i) \in \mathbb{N}$.
2. Instances of coherence types are either paratactic (symmetric, undirected) or hypotactic (asymmetric, directed). This is reflected by a classification $f_c\colon K \to \{\mathit{para}, \mathit{hypo}\}$. □

It is clear that this definition is (apart from its incompleteness) far too simple regarding the complexity of coherence relations, their relational type, (partly dynamic) arity, and roles of their arguments. Although a hypergraph-based definition would thus be more adequate, we keep this simplified version since it is sufficient to outline the general building blocks of SFHTs. In order to introduce coherence relations as instances of coherence types, we need to auxiliary definitions, which introduce the important concept of a text base:

**Definition 6** *Segmentation.* Let $C = \{x_1, \ldots, x_n\}$ be a corpus of natural language texts and $\mathbf{A}$ an algorithm, which completely segments texts $x \in C$ into their non-overlapping, hierarchically organized components. $\mathbf{A}$ induces a segmentation function $f_{\mathbf{A}}\colon C \to \{f_{\mathbf{A}}(x) \mid x \in C\}$, which maps each text $x \in C$ onto the finite set of its $\mathbf{A}$-based segments $f_{\mathbf{A}}(x) \ni x$. In order to impose a structure over $f_{\mathbf{A}}(x)$, we suppose that $\mathbf{A}$ segments texts $x \in C$ into trees $I(x) = \langle f_{\mathbf{A}}(x), E, x\rangle$ with root $x$, where the leafs of $I(x)$ (e.g. the lexical tokens of $x$) correspond to the elementary segments of $x$ according to $\mathbf{A}$. $I(x)$ is called *segmentation* or *integration hierarchy* of $x$. □

**Definition 7** *Text Base.* Let $C = \{x_1, \ldots, x_n\}$ be a text corpus with segmentations $I(x_1), \ldots, I(x_n)$ according to algorithm $\mathbf{A}$. The *text base TB* induced by $\mathbf{A}$ over $C$ is a forrest $TB = \langle Y, I\rangle$, where $Y = \bigcup_{x \in C} f_{\mathbf{A}}(x)$ and $I = \bigcup_{x \in C, I(x) = \langle V, E, x\rangle} E$. □

With the help of the concept of a text base coherence relations can now be defined as relations of text segments. This is done by analogy with the distinction of *relation formats* (coherence types), *tuples* (coherence relations) and *relations* (coherence sets as sets of tuples of the same format) in the relational database model:

**Definition 8** *Coherence Relations and Coherence Sets.* Let $TB = \langle Y, I\rangle$ be a text base induced by a segmentation algorithm $\mathbf{A}$ over a corpus $C$. A *coherence relation* $(y_1, \ldots, y_{s_i})$ of type $k_i \in K$ with signature $s_i$ is an element of $Y^{s_i}$. In case of binary types $k$ we write $y_i \Leftrightarrow_k y_j$ for symmetric and $y_i \Rightarrow_k y_j$ for asymmetric

[9] $K$ may be identified, for example, with the set of coherence relations described by [Martin, 1992].

relations, respectively. If all components of a coherence relation belong to the same text $x \in C$, it is called *intratextual*, else it is called *intertextual*. A *coherence set* of type $k_i \in K$ with signature $s_i$ is a subset of $Y^{s_i}$. A *fuzzy coherence set* of type $k_i \in K$ with signature $s_i$ is a fuzzy set over $Y^{s_i}$. □

Once more, the reference to fuzzy sets in [Def. 8] serves to abstract from the concrete concept of informational uncertainty *without* excluding probabilistic weightings of coherence relations. In order to give now an example of the mapping of linguistic realization patterns onto genres and registers, we concentrate on the coherence type of unsystematic lexical cohesion as described by [Halliday and Hasan, 1976] as a means for linguistic manifestation of register networking:

**Definition 9** *Lexical Realization and Lexical Field.* Let $RL = \langle V, \mathcal{E}, R\rangle$ be a register layer derived from a sequence $\langle V_0 \cup T, \mathcal{E}_0\rangle, \ldots, \langle V_n, \mathcal{E}_n\rangle$ of hypergraphs according to [Def. 4]. Let further $W$ be a set of lexical units. A lexical realization of a topic vertex $v \in V_0$ is a fuzzy set $\mathcal{W}_v$ over $W$. Let $r \in R$ be a register with elementary topic vertices $V_0(r) \subseteq V_0$. The lexical realization $\mathcal{W}_r$ of $r$ is the union of the lexical realizations of its elementary topic vertices and is called *lexical field* of $r$: $\mathcal{W}_r = \bigcup_{v \in V_0(r)} \mathcal{W}_v$. □

The register layer can now be augmented by lexical fields as a kind of linguistic realization pattern:

**Definition 10** *Augmented Register Layer.* Let $RL = \langle V, \mathcal{E}, \{\langle V_i, \mathcal{E}_i\rangle \,|\, V_i \subseteq V, \mathcal{E}_i \subseteq \mathcal{E}, i \in \mathfrak{I}\}\rangle$ be a register layer according to [Def. 4]. An *augmented register layer* is a tuple $RL' = \langle V, \mathcal{E}, \{\langle V_i, \mathcal{E}_i, \mathcal{W}_{\langle V_i, \mathcal{E}_i\rangle}\rangle \,|\, V_i \subseteq V, \mathcal{E}_i \subseteq \mathcal{E}, i \in \mathfrak{I}\}\rangle$, where $\mathcal{W}_{\langle V_i, \mathcal{E}_i\rangle}$ is the lexical field of $\langle V_i, \mathcal{E}_i\rangle$ according to [Def. 9]. □

**Definition 11** *Lexical Text Categorization.* Let $TB = \langle Y, I\rangle$ be a text base according to [Def. 7] and $RL = \langle V, \mathcal{E}, R\rangle$ a register layer according to [Def. 10]. A *text categorization* is a function $\varphi\colon Y \rightarrow \mathbb{P}(R)$, which maps each text segment $y \in Y$ onto a set of registers according to a comparison of its lexical organization with the lexical realization $\mathcal{W}_r$ of these registers. $\mathbb{P}(R)$ is the power set of $R$. □

The connection of contextual grounding, coherence relations, and text linkage, to which this definition refers, can be exemplified as follows: Suppose a register of unequal tenor and non-interactive mode concerning *automobile industry* (field)—a register typically realized by newspaper articles. In this example, a lexical field realizing the register can be assumed, which contains elements like *car*, *automobile*, *serial production*, etc. Furthermore, a related register can be supposed, e.g. a mode- and tenor-equal register concerning *stock market*, whose lexical realization comprises sense or associatively related lexical units, e.g. *share*

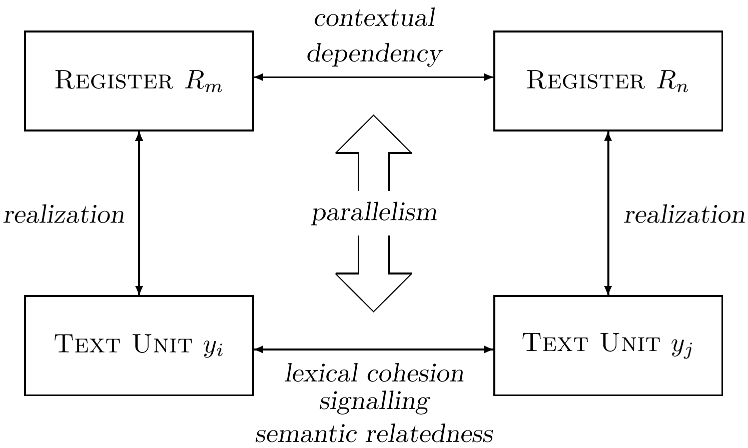


**Figure 2:** The parallelism of contextual and textual dependencies.

*holder value*, *interest rate*, etc. On this background, it is easy to identify two texts $x, y$ realizing these registers and cohering on the basis of the lexical cohesion of their vocabulary. In this sense, the intertextual coherence relation $(x, y)$ realizes as an instance of the coherence type of lexical cohesion a dependency relation of two registers *by means of instantiating elements of the registers' lexical fields.* As a consequence, as visualized in [Fig. 2], the coherence relation is paralleled and thus becomes interpretable by the contextual dependency of the registers in question. In the following, links are used to *manifest* this twofold grounding of text relations.

### 3.1 Systemic Functional Links

In SFHTs, links are seen to be conditional on linguistic and contextual grounding in the sense that the same link may be supported by (and thus being interpretable according to) different linguistic and contextual resources. As a kind of digital sign, links serve as hypertextual manifestations of intra- and intertextual coherence relations, which on their turn may realize dependencies of contextual units (e.g. of constituents of schematic structure). Links characterized along these lines will be called *systemic functional links* (SFLinks). According to the general conception that SFLinks are signs, content and expression plane have to be distinguished as their organizational planes: Suppose a binary coherence relation $(y_i, y_j)$ of type $k_i \in K$ to be manifested by a SFLink $l$, where text span $y_i$ hypotactically depends on text span $y_j$ (that is the interpretation of $y_i$ in the present text depends on the interpretation of $y_j$ in the same or another text; in order to keep the formalism simple we suppose from now on that coherence types have binary signatures leaving the formalization of more complex coherence relations and their hypertextual manifestations to future work):

1. The *expression plane* of $l$ is given by the segments $y_i, y_j$ and the anchor $a$ of $l$ as part of $y_i$ (possibly added by $l$'s anchor in $y_j$). If $l$ is used to manifests several coherence relations, they need to connect the same segments. Furthermore, depending on the semantics of the coherence types $l$ manifests, it is either paratactic or hypotactic, i.e. navigatable in both or only one direction.
2. The *content plane* of $l$ is defined by the linguistic and contextual configuration it instantiates:
   - *Texture plane:* The types of coherence relations manifested by $l$ define its *textual support*, which specifies how the link target has to be interpreted from the perspective of its source, and vice versa.
   - *Context plane:* A coherence relation is not just an instance of a coherence type. Which type is instantiated where, in which order and how often depends on context. Thus, SFLinks are only properly interpreted on the background of relations of context units, which parallel coherence relations and their manifestations by SFLinks. These contextual relations define a link's *contextual support*.

Concentrating on binary, hypotactic links as manifestations of lexically cohesive texts, this can be put into more formal terms:

**Definition 12** Let $RL = \langle V, \mathcal{E}, R\rangle$ be a register layer according to [Def. 10], and $TL$ a texture layer with the set of coherence types $K \ni u$, where $u$ is the type of unsystematic lexical cohesion. Let further $TB = \langle Y, I\rangle$ be a text base induced by a segmentation algorithm $\mathbf{A}$ over corpus $C$. Let finally $\varphi\colon Y \to \mathbb{P}(R)$ be a text categorization according to [Def. 11]. A *binary lexically cohesive link with registerial support* is a tuple $l = \langle y_i, y_j, a, \{u\}, \{r_m, r_n\}\rangle$, where the *source* span $y_i \in Y$ is linked via anchor $a \in f_{\mathbf{A}}(y_i)$ with the *target* span $y_j \in Y$ so that $y_i \Rightarrow_u y_j$ is a coherence relation of type $u$ and $\mathcal{E}$ contains an edge which connects the registers $r_m, r_n \in R$, where $r_m \in \varphi(y_i), r_n \in \varphi(y_j)$. $\{r_m, r_n\}$ is called *registerial support* of $l$. □

Whereas the *texture plane* of a link restricts *how* it has to be interpreted, it is the *context plane*, which helps to determine *what* relation the link actually manifests. Obviously, the planes of a SFLink are based on a sequence of realizations: dependencies of contextual units are realized by coherence relations as instances of coherence types, which on their turn are manifested by SFLinks and their markers. Thus, SFLinks are a kind of *digitalized sign*: they are hypertextual, interactivity providing realizations of coherence relations. The recursive realization

$$\begin{tikzcd}
Co_i \arrow[r, "\Rrightarrow"] \arrow[d, "\searrow"'] & Co_j \arrow[d, "\searrow"] \\
x_m \arrow[r, "\Rightarrow"] \arrow[d, "\searrow"'] & x_n \arrow[d, "\searrow"] \\
x_m \arrow[r, "\rightarrow"] & x_n
\end{tikzcd}$$

**Figure 3:** Recursive realization of contextual, text semantic and textual units.

leading to SFLinks is shown in [Fig. 3], where $\searrow$ symbolizes realization, $x_m, x_n$ are text segments linked by a SFLink ($\rightarrow$) and related by a coherence relation ($\Rightarrow$), respectively; finally, $Co_i$, $Co_j$ are contextual units related by a dependency relation ($\Rrightarrow$). Thus, a fundamental characteristic of a SFLink $x_m \rightarrow x_n$ is that it parallels a coherence relation $x_m \Rightarrow x_n$, which on its part parallels a dependency relation $Co_i \Rrightarrow Co_j$ of genres, stages, registers, or register variables.

### 3.2 Systemic Functional Paths

Whereas the criteria of textual and contextual support serve to provide interpretability of binary links and thus to avoid negative effects of text linkage, it is the criterion of structure formation, which serves to avoid negative effects of purely associative link chaining. These negative effects are obvious when looking at intransitive similarity relations of textual units to be used as a source of associative text linkage: In case of two pairs of texts $x, y$ and $y, z$, where each is more lexically similar than the pair of texts $x$ and $z$, there is a risk to produce the thematically diversifying path $(x, y, z)$. The less the degree of similarity of $x$ and $z$ compared to the similarity degree of the texts $x, y$, and $y, z$, respectively, the higher the risk of a thematic diversification, *already after two links*. It is obvious that chains of links of this sort, which result from a lack of context-sensitive, top-down control of text linkage, produce the well known problem of disorientation in hypertext.

In [Mehler, 2002a, Mehler, 2002b] it is shown, how to reduce this risk by retarding topic changes in paths of interlinked documents. So called *cohesion trees* are proposed, which make corpora traversable by means of hierarchies, whose branches code different thematic aspects primed by the same temporary root. Coherence trees are based on three (ideal) criteria for structuring corpora:

1. *Thematic progression* of texts have to be modelled by paths, whereby latent topic changes occur as the path length grows.

$$\begin{tikzcd}
Co_{i_1} \arrow[r, "\Rrightarrow"] \arrow[d, "\searrow"'] & Co_{i_2} \arrow[r, dotted, no head, "\Rrightarrow"] \arrow[d, "\searrow"] & Co_{i_{n-1}} \arrow[r, "\Rrightarrow"] \arrow[d, "\searrow"'] & Co_{i_n} \arrow[d, "\searrow"] \\
x_{j_1} \arrow[r, "\Rightarrow"] \arrow[d, "\searrow"'] & x_{j_2} \arrow[r, dotted, no head, "\Rightarrow"] \arrow[d, "\searrow"] & x_{j_{n-1}} \arrow[r, "\Rightarrow"] \arrow[d, "\searrow"'] & x_{j_n} \arrow[d, "\searrow"] \\
x_{j_1} \arrow[r, "\rightarrow"] & x_{j_2} \arrow[r, dotted, no head, "\rightarrow"] & x_{j_{n-1}} \arrow[r, "\rightarrow"] & x_{j_n}
\end{tikzcd}$$

Figure 4: The parallelism of text linkage ($\rightarrow$), coherence relations ($\Rightarrow$), and systemic functional progressions as sequels of contextual dependencies ($\Rrightarrow$) on the level of systemic functional paths.

2. *Thematic branching:* Thematically ambiguous texts have to be modelled as branching nodes whose outgoing paths represent different thematic aspects connoted by the ambiguous text node in question.

3. *Interactivity:* Any node of a cohesion tree can be chosen at runtime to be the root of a new tree for traversing the same text corpus from the perspective of the topic connoted by this text.

In spite of these building principles cohesion trees still lack any context driven, top-down control of topic tracking: In case of a text $z \in C$ to be inserted into a cohesion tree $CT(x)$ of a text $x \in C$, the algorithm for generating cohesion trees does not choose those text $y \in C$ already inserted into $CT(x)$, which is most similar to $z$, but the end vertex of those path $\mathcal{P}$ in $CT(x)$ starting with $x$, which minimizes loss of cohesion, when $z$ is attached to it, as $z$'s predecessor in $CT(x)$. This bottom-up procedure (leaving out any top-down control by means of preestablished context representations above the level of texts) still runs the risk to order texts close to each other, which—in spite of having similar lexical organizations—diverge thematically.

In order to overcome this risk, the concept of *systemic functional path* (SFPath) is proposed, which extends the idea of textual and contextual support onto the level of paths by using SFLinks as their constituents. This is exemplified in [Fig. 4], where SFLinks enter into a chain of links, which as a whole does not only parallel a chain of coherence relations ($\Rightarrow$), but also a henceforth called systemic functional progression (SFProgression) of contextual (registerial/generic) units ($\Rrightarrow$), where these two chains as a whole give textual and contextual support to the SFPath. We formalize this by extending the simplified version of a SFLink according to [Def. 12]:

**Definition 13** Let $RL = \langle V, \mathcal{E}, R\rangle$ be a register layer according to [Def. 10]. A *systemic functional path with registerial support* is a sequence $p = (y_{i_1}, \ldots, y_{i_n})$

of text spans, where for all $j \in \{1, n-1\}$ $(y_{i_j}, y_{i_{j+1}})$ is a SFLink according to [Def. 12] with registerial support $\{r_{m_j}, r_{m_{j+1}}\}$, whereby the union of registerial supports $U = \{r_{m_1}, \ldots, r_{m_n}\}$ is the vertex set of a path in $\mathcal{E}$. $U$ is called registerial support of $p$. $\square$

What does it mean to call a link or path 'systemic functional'? It means that they serve to manifest dependencies of contextual units, of genres and registers, and thus of social-semiotic entities which restrict as well as result of countless many communication acts of a speech community. In this sense, not only the formation of links, but also of paths is seen to be underpinned by dependencies, or as in case of SFPaths: of SFProgressions of contextual units. This is exemplified in [Fig. 5], where—according to the layered architecture of SFHTs—link base and context (i.e. genre and register) layer(s) are separated. In case of a SFLink $l$ between text units $x_i, x_j$, structure formation is not exhausted by specifying the type of coherence relation manifested by $l$, whose interpretation is restricted by a dependency relation $d$ of such units $Co_m, Co_n$. Rather, the participation of $l$ in the formation of path $\mathcal{P}_1$ (as an alternative to $\mathcal{P}_2$) is evaluated on the background of a corresponding progression of contextual units, symbolized by $\Psi$, into which the dependency relation $d$ of $Co_m, Co_n$ enters. It this progression $\Psi$ (e.g. a sequel of stages defining a genre or a sequence of related topics), which as a whole serves as a source for interpreting the sequel of links forming $\mathcal{P}_1$. In this sense, SFPaths are specified as a special class of signs, whose content plane is—comparable to SFLinks—described with respect to their textual and contextual support. But other than SFLinks, SFPaths are complex hypertextual signs, whose contextual support does not consist of single contextual dependency relations, but of progressions of context units.

On this background, linguistic and contextual support as well as structure formation appear to be fundamentals of SFHTs:

1. *Support:* SFLinks are used as manifestations of coherence relations, which on their turn realise dependencies of contextual units, and
2. *Structure Formation:* SFLinks enter into SFPaths, which as a whole are contextually supported by progressions of contextual units.
3. Finally, SFHTs appear as supersigns, whose immediate constituents are SFPaths (as well as composite nodes [Halasz, 1988]).

SFPaths allow to implement the concept of *backward* and *forward navigation* (see [Thüring et al., 1991] for these concepts): Any position in a SFPath is interpretable with respect to the predecessor nodes it continues (interprets, evaluates, summarizes) and successor nodes, whose interpretation it primes, prepares, restricts, etc. This allows the reader to ask for the derivation/continuation of the actual node $x$ according to the guideline of the progression to which $x$

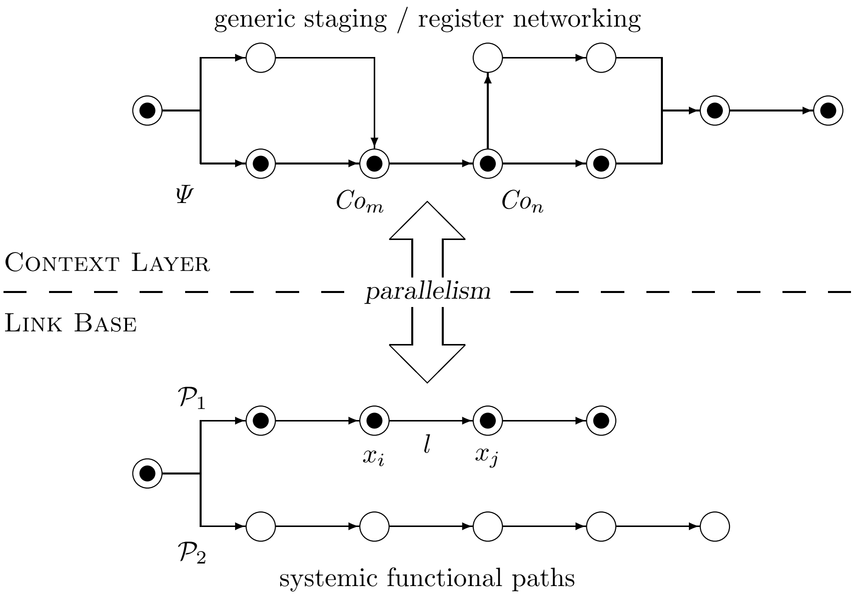


**Figure 5:** The parallelism of context layer and link base

belongs, where navigational questions (*Where do I come from?* or *Where to go next?*) are answered in systemic functional terms, that is, with respect to the manifestation of generic staging or registerial networking of the node, link or path to be evaluated.

## 4 An Example

In order to exemplify SFHTs and their building blocks, suppose a corpus of newspaper articles with texts about the election of the Lower House of German Parliament (field). In this example of a homogeneous field we expect texts of varying (sub-)genres (of the genre of newspaper articles): general reports (on the course of the election, its results, reactions, etc.), texts integrating diagrams/tables (concerning election prognoses, results, voter turnout, etc.), leading articles and commentaries (e.g. of political scientists), interviews, essays as part of the feature pages, portraits and biographies (of persons involved), press review, letters to the editor, etc. Further, the field in question can be seen to be structured into many sub-fields concerning the election campaign, election results (in the Federal Republic compared with the results in the single federal states), the voter turnout, changes compared with preceding elections, reactions of the government/opposition/employers' association/labor union/allied countries/the stock market, analyses of the results of single parties, difference between West and Eastern Germany, etc. Furthermore, many related fields can be enumerated which are easily accessible from the field of election. This concerns the fields of economic development, home policy, foreign policy as well as preceding elections, the history of the democratic system in Germany, etc.

The question arises, how to explore this diversity in hypertext. The specific answer SFHTs give is to control text linkage as well as chaining of links and their clustering (in case of composite nodes) by register networking and generic staging, i.e. by systemic functional links and paths preserving—as far as possible—registerial and generic dependencies. In the present example this means that SFPaths are generated in which texts have a higher probability to be linked when dealing with the same (sub-)register and instantiating the same or subsequent genres (stages) than texts varying according these dimensions. In paths of this kind, letters to the editor *are* linked with the article they comment, portraits of politicians *are* linked with their interviews, reports on reactions from abroad do *not* precede, but succeed reports on the election results they deal with, etc. As a result, a network of SFPaths is generated in which links are used to manifest registerial and/or generic turns, whereas paths manifest the thematic/chronological/generic ordering of the corpus. Thereby, branching nodes manifest "ambiguous" texts (with respect to the genres and/or registers they instantiate) as starting points of different, but genre and/or register-related paths, whereby forking branches may converge by being linked with the same summary/concluding texts. In this example, texts are not simply linked because they are judged to be similar according to some criterion of lexical similarity, but because their linkage manifests a generic/registerial turn *as part of a progression of coherent turns, each of which is manifested by a SFLink*.

What—in contrast to this example—would those approaches produce, which operate in the framework of associative text linkage (as for examples approaches based on Salton's vector space model)? Since they rely on numerical measures exploiting similarities of lexical organization, they would preferably link those texts which share to a higher degree more important words (where importance is measured for example in terms of the *tfidf*-scheme). Thereby, neither context models, nor restrictions of structure formation (beyond numerical limits concerning the number of links) are applied. As a consequence, texts dealing with elections have a high chance to be interrelated irrespective of the concrete aspect of the concrete election they deal with and the actual stage or genre they instantiate. Furthermore, the aspect according to which texts are linked may change from link to link without being explicitly marked. This interference of interpretability produces negative effects especially with respect to structure formation: because of the intransitivity of similarity relations resulting from the measures used, registerial and generic diversifications may already occur after view links, whereby barely interpretable networks emerge on the level of the whole hypertext. More coherent structures are produced in the framework of *topic tracking* [Carthy and Smeaton, 2000], e.g. sequences of thematically homogeneous, chronologically ordered newspaper articles. But this approach still relies on the vector space model and similarity measures derived from it. Thus, register-controlled branching as well as generic staging are still left out. SFHTs serve to break with these approaches by means of focusing on the text linguistic basis of *text* linkage in hyper*text*.

## 5 Conclusions

A new architecture for hypertext authoring is described based on a linguistic theory. This model does not only focus on higher level link structures, but on a context model as the proper basis of text linkage, where the concept of SFLink is introduced in order to provide text linguistic interpretability of links, whereas the concept of a SFPath is introduced in order to provide interpretability above the level of single links. SFLinks and SFPaths serve for a theoretical linguistic underpinning of hypertext, or more general: for reconstructing hypertexts as text linguistic objects based on a proper computational linguistic framework.

## References

[Agosti et al., 1997] Agosti, M., Crestani, F., and Melucci, M. (1997). On the use of information retrieval techniques for the automatic construction of hypertext. *Information Processing & Management*, 33(2):133–144.

[Agosti and Smeaton, 1996] Agosti, M. and Smeaton, A. F. (1996). *Information Retrieval and Hypertext*. Kluwer, Boston.

[Allan, 1997] Allan, J. (1997). Building hypertext using information retrieval. *Information Processing & Management*, 33(2):145–159.

[Baas, 1994] Baas, N. A. (1994). Emergence, hierarchies, and hyperstructures. In Langton, C. G., editor, *Artificial Life III, SFI Studies in the Sciences of Complexity*, pages 515–537. Addison-Wesley.

[Barwise and Perry, 1983] Barwise, J. and Perry, J. (1983). *Situations and Attitudes*. MIT Press, Cambridge.

[Biber, 1991] Biber, D. (1991). *Variation Across Speech and Writing*. Cambridge University Press, Cambridge.

[Carthy and Smeaton, 2000] Carthy, J. and Smeaton, A. F. (2000). The design of a topic tracking system. In *Proceedings of the 22nd Annual Colloquium on Information Retrieval Research*, Cambridge, England. The Information Retrieval Specialist Group of the British Computer Society.

[Chen, 1997] Chen, C. (1997). Structuring and visualising the www by generalised similarity analysis. In Bernstein, M., Carr, L., and Østerbye, K., editors, *Proceedings of the Eighth ACM Conference on Hypertext and Hypermedia*, pages 177–186, New York. ACM.

[Dalamagas and Dunlop, 1997] Dalamagas, T. and Dunlop, M. D. (1997). Automatic construction of news hypertext. In Fuhr, N., editor, *Hypertext – Information Retrieval – Multimedia. Proceedings of the HIM '97*, pages 265–278. Universitätsverlag, Konstanz.

[El-Beltagy et al., 2001] El-Beltagy, S. R., Hall, W., Roure, D. D., and Carr, L. (2001). Linking in context. *Journal of Digital Information*, 2(3).

[Fellbaum, 1998] Fellbaum, C., editor (1998). *WordNet : An Electronic Lexical Database*. MIT Press, Cambridge.

[Ferret, 2002] Ferret, O. (2002). Using collocations for topic segmentation and link detection. In *Proceedings of the 19th International Conference on Computational Lingjuistics, COLING 2002, Taipei*, pages 260–266. Morgan Kaufmann.

[Green, 1998] Green, S. J. (1998). Automated link generation: Can we do better than term repetition? *Computer Networks and ISDN Systems*, 30(1-7):75–84.

[Halasz, 1988] Halasz, F. G. (1988). Reflections on notecards: Seven issues for the next generation of hypermedia systems. *Communications of the ACM*, 31(7):836–852.

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

[Halliday and Hasan, 1976] Halliday, M. A. K. and Hasan, R. (1976). *Cohesion in English*. Longman, London.

[Halliday and Hasan, 1989] Halliday, M. A. K. and Hasan, R. (1989). *Language, Context, and Text: Aspects of Language in a Socialsemiotic Perspective*. Oxford University Press, Oxford.

[Knott and Sanders, 1998] Knott, A. and Sanders, T. (1998). The classification of coherence relations and their linguistic markers: An exploration of two languages. *Journal of Pragmatics*, 30:135–175.

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

[Mann and Thompson, 1988] Mann, W. C. and Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8:243–281.

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

[Mehler, 2002a] Mehler, A. (2002a). Hierarchical analysis of text similarity data. *Künstliche Intelligenz*, 2:12–16.

[Mehler, 2002b] Mehler, A. (2002b). Hierarchical orderings of textual units. In *Proceedings of the 19th International Conference on Computational Lingjuistics, COLING 2002, Taipei*, pages 646–652. Morgan Kaufmann.

[Mehler, 2002c] Mehler, A. (2002c). Textbedeutungsrekonstruktion. Grundzüge einer Architektur zur Modellierung der Bedeutungen von Texten. In Pohl, I., editor, *Prozesse der Bedeutungskonstruktion*, pages 445–486. Peter Lang, Frankfurt a.M.

[Salton et al., 1994] Salton, G., Allan, J., and Buckley, C. (1994). Automatic structuring and retrieval of large text files. *Communications of the ACM*, 37(2):97–108.

[Thüring et al., 1991] Thüring, M., Haake, J. M., and Hannemann, J. (1991). What's eliza doing in the chinese room? Incoherent hyperdocuments - and how to avoid them. In *Proceedings of the Third AM Conference on Hypertext*, pages 161–177. ACM.

[van Dijk and Kintsch, 1983] van Dijk, T. A. and Kintsch, W. (1983). *Strategies of Discourse Comprehension*. Academic Press, New York [u.a.].

[Ventola, 1987] Ventola, E. (1987). *The Structure of Social Interaction: a Systemic Approach to the Semiotics of Service Encounters*. Pinter, London.

[Widhalm and Mück, 2001] Widhalm, R. and Mück, T. (2001). *Topic Maps. Semantische Suche im Internet*. Springer, Berlin.

[Wilkinson and Smeaton, 1999] Wilkinson, R. and Smeaton, A. F. (1999). Automatic link generation. *ACM Computing Surveys*, 31(4).

[Zellweger, 1989] Zellweger, P. T. (1989). Scripted documents. A hypermedia path mechanism. In *Proceedings of the Second ACM Conference on Hypertext*, pages 1–14, New York. ACM.