Identification of Topic-Focus Chains (work in progress)

Ivana Korbayová-Kruijffová[1] *and Geert-Jan Kruijff*

Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University
Malostranské náměstí 25, 118 00 Praha 1, Czech Republic; *{korbay, gj}@ufal.mff.cuni.cz*

Abstract

A coherent discourse forms a continuum consisting of interrelated text units. This paper draws on the assumption that a text unit is articulated into topic and focus, and that the topic of a text unit is linked to the preceding context by a TFA *connection. A succession of* TFA *connections throughout a discourse forms a TFChain the structure of which can be captured by grammar. Underlying the* TFA *connections are contextual relationships of identity, similarity or contiguity at various levels of text unit representation, including its outer shape, underlying structure and referential contents. Identification of a TFChain is seen to guide discourse interpretation.*

1. Introduction

It can be assumed that a coherent discourse consists of interrelated text units, because coherent discourses tend to be about related things, or the assertions they express about different things tend to be related in some way[2]. A text unit in this treatment corresponds roughly to a simple clause (a clause containing one verb and its modifiers). Following the approach advocated by the Modern Prague School, we assume that every text unit is articulated into Topic and Focus, although the topic may be empty in some cases (cf. Topic-focus articulation [Sgall et al. 1986]).

In the present paper, we want to specify how a text unit relates to its preceding context in terms of a link between its topic and the topic, the focus or the whole of some preceding text unit. The relation between two text units will be called a *TFA connection*. Our notion of TFA connections has been inspired mainly by the works of Weil [1887], Daneš [1979], Hajičová & Sgall [1984] and Daneš [1985].

A sequence of TFA connections throughout a discourse will be called *Topic-Focus Chain* (TFChain) henceforth. A TFChain reflects the flow of information (contents) in a discourse. We therefore regard identification of a TFChain as crucial to comprehension of the information conveyed by the text, as far as the referential function of language is concerned.

Underlying the TFA connections are anaphoric relations we call *contextual relationships*. We distinguish contextual relationships at several levels of interpretation and representation. The contextual relationships we distinguish correspond to *identity* or various forms of *similarity* or *contiguity*.

Our aim in the present paper is to specify TFA connections, TFChains and the contextual relationships underlying them, and to show how TFChains can be determined during discourse analysis. The structure of this paper is as follows:

We overview the work of Daneš which has been our main source of inspiration in Section 2. Then, the notions of TFA connection and TFChain are introduced in Section 3, where we also present a classification of TFA connections and a skeleton of TFChain grammar. The contextual relationships underlying TFA connections are discussed in Section 4. We describe the levels of representation at which they are discerned, classify their types and specify correlations between contextual relationships and TFA connections. We complete our presentation of TFChains by proposing an algorithm for discourse interpretation that has TFChain identification as its backbone, in Section 5. Section 6 contains our conclusions and suggestions for further research.

---

[1]The author has been receiving support from the Research Support Scheme of the Open Society Institute, grant no. 66/64.

[2]As far as we are aware, the notion of "articulation of connected discourse" was first introduced by Weil [1887]. As A. Scaglione puts it in the introduction to a translation of Weil's article, "Generally, Weil's system could be characterized as a contextualist one: for him the specific arrangement of the parts of a discourse cannot be comprehended without knowing the total meaning of the clause, i.e., what is in it and what precedes it. The sequence of clauses and sentences forms a *continuum*. We have therefore, in a well-constructed discourse, a *marche des idees* which can proceed either in a parallel pattern, by *reprise de la notion initiale*, or in a progressive (somewhat 'chiastic') pattern, by picking up 'the goal' (*le but*) of the last sentence or clause."

2. Earlier work: thematic sequences

We have drawn substantially on the proposals in [Daneš 1979] and [Daneš 1985]. Daneš's theory as sumes that an utterance (U) is structured into two parts, being the *theme* (T), i.e., that what is talked about, and the *rheme* (R), i.e., that what is said about the theme. He distinguishes various types of *th ematic sequences* according to what is being thematized in the current utterance, whether the theme of the current utterance is a repetition or a derivation of what it thematizes, and whether what is bein g thematized in the current utterance occurs in the immediately preceding utterance (contact versus distant thematic sequences, cf. also Section 3.1).

With respect to what is being thematized, Daneš distinguishes the following basic types of *contact t hematic sequences*[3]:

1. thematization of the preceding rheme
   i. theme is a repetition of the preceding rheme: $T_{i+1}=R_i$[4]
   ii. theme is a derivation from the preceding rheme: $T_{i+1}<R_i$
2. thematization of the preceding utterance: $T_{i+1}=U_i$
3. continuous theme:
   i. the theme is a repetition of the preceding theme: $T_{i+1}=T_i$
   ii. the theme is a derivation from the preceding theme: $T_{i+1}<T_i$
4. the theme is derived from a hyper-theme, i.e., the theme of a super-ordinate text unit, e.g., a text paragraph: $T_{i+1}<T^*$
5. the theme expresses a summarization of several preceding utterances, U_i through U_j, $i<j$, constitu ting a continuous segment: $T_{i+1}=I(i,j)$

U_{i+1}, $i>1$, denotes the utterance currently evaluated. The thematic sequence linking U_{i+1} to the prece ding discourse is evaluated by looking from the currently processed utterance backwards.

The example below illustrates the various types of thematic sequences:[5]

Example 1:

0. *Národní muzeum* stojí na Václavském náměstí.
   [*The National museum* stands on the Venceslaus square.]
1. (a) *Toto náměstí* je jedním z nejrušnějších míst Prahy. $T_{1a}=R_0$
   [*This square* is one of the most busy places in Prague.]
   (b) *Horní část tohoto velkého prostranství* se tak dostalo krásné dominanty $T_{1b}<R_0$
   [*The top part of this large area* has thus received a nice dominant.]
2. *Tato skutečnost* je známa snad každému návštěvníkovi Prahy. $T_2=U_0$
   [*This fact* is known by perhaps every visitor of Prague.]
3. (a) Je *to* velmi památná budova. $T_{3a}=T_0$
   [*It* is a very memorial building.]
   (b) *Sbírky Národního muzea* představují významnou národní kulturní hodnotu. $T_{3b}<T_0$
   [*The collections of the National museum* represent an important national cultural value.]
4. *Jiná mimořádně významná pražská budova, Národní divadlo*, je umístěna na Smetanově nábřeží.
   $T_4<T^*$, where T* is 'významné pražské budovy' (important Prague buildings')
   [*Another remarkably important Prague building, the National theatre*, is situated on the Smetana embankment.]

Daneš also distinguishes simple, conjoined and condensed utterances, with respect to their theme-rh eme structure. We shall discuss this subject in Section 0.

Daneš calls the couples of related elements *isotopic pairs* and discusses the *isotopic relationships* be tween the related elements. He distinguishes relationships of *referential identity (co-reference)* and r elationships of *referential non-identity* (various types of semantic relatedness). In our understanding

---

[3]If we liken Daneš's notions of *theme* and *rheme* with Weil's *la notion initial* and *le but,* respectively, we can say that Daneš's thematization of the rheme corresponds to Weil's progressive pattern, whereas Daneš's continuous theme corresponds to Weil's parallel pattern. Daneš's thematic sequences with a summarizing theme or a theme derived from a hyper-theme do not have a straightforward equivalent in Weil's approach. The summarizing theme could be perhaps subsumed under the progressive pattern, and the hyper-theme derivation under the parallel pattern.

[4]The notation employed here is from [Daneš 1985].

[5]The example is adopted from [Daneš 1985]. Italics are ours and depict the themes according to Daneš. Translation into English is also ours and is almost literal.

of his approach, referential identity corresponds to repetition and referential non-identity corresponds to derivation in his classification as reproduced above. We shall return to the subject of such relationships later on (Section 4.2). Let us now remark on Daneš's classification of thematic sequences, in order to motivate .

In our opinion, Daneš's types 2 and 5 can be replaced by one type where the theme is a summarization of one or more preceding utterances constituting a *continuous segment*: $T_{i+1}=I(j,i)$, $j \leq i$. If one wants to employ this type, one has to specify what is meant by a continuous segment. This more general notion of summarized theme will then be more accurate, because a single complex idea can be expressed in one utterance or developed over several utterances, as the following example demonstrates.

Example 2

(The text had to be abbreviated due to lack of space. In our interpretation of the text, the anaphoric expression 'tyto úvahy' ('these considerations') in 2 expresses a summarization of the ideas introduced by the if-clauses in the preceding paragraph.)

1. (a paragraph on a project of the Apple company concerning special radio-communication systems for speech and data transmission, including several if-clauses about some issues of such systems)
2. *Tyto úvahy* jsou ale patrně pouze akademické, neboť po ohlášení tohoto plánu se ozvalo v odpověď vícenásobné protestní kdákání několika organizací, ...
   [*These considerations* are however apparently only academic, because after the announcement of this project, multiple protest quacking of several organizations, ..., has sounded in response.]

It is also important to note that the thematic sequence of type 4 as specified by Daneš differs in nature from the other types. His example (reproduced in Example 1.4 above) demonstrates the case where the hyper-theme is explicitly mentioned. But in this case, one could also say that the theme 'jiná mimořádně výnamná pražská budova, Národní divadlo', is derived from the theme 'Národní muzeum', on the basis of set-member relationship: 'Národní muzeum' is included among the important Prague buildings, 'významné pražské budovy', and so is 'Národní divadlo'. 'Národní muzeum' and 'Národní divadlo' can be therefore considered as co-hyposememes (referring to semantically co-ordinated entities).

If the hyper-theme is not explicitly mentioned in the discourse[6], the recipient of the discourse may infer it as a "unifying concept". When identifying a hyper-theme not explicitly mentioned, it is first necessary to identify a relation between themes, e.g., co-hyposemy, and on this ground it can be inferred that the common super-ordinate concept constitutes a hyper-theme. We prefer to consider derivation from hyper-theme as a type of thematic sequence belonging to a different level of interpretation, because it involves taking into account implicit relations. Consider the following discourse, which is a modification of Daneš's example:

Example 3

0. *Národní muzeum* stojí na Václavském náměstí,
   [*The National museum* stands on the Venceslaus square,]
1. kdežto *Národní divadlo* je umístěno na Smetanově nábřeží.
   [whilst *the National theatre* is situated on the Smetana embankment.]

The themes of the two clauses are related as siblings, since they are both buildings 'of the nation'. The rhemes of the clauses are also related, each of them predicating a location at a different place in Prague. So, the themes have even more in common: they are both located in Prague. Furthermore, the two clauses are co-ordinated. Therefore, one can note the parallel structure. By abstraction, one can infer a common hyper-theme embracing these two themes and thus connection between the themes of the two clauses, namely "national buildings in Prague". The fact that the hyper-theme has been inferred rather than explicitly mentioned in the text results, however, in a different status of the connection between the clauses.

---

[6]We are not sure whether Daneš had in mind hyper-themes both explicitly mentioned and implicit.

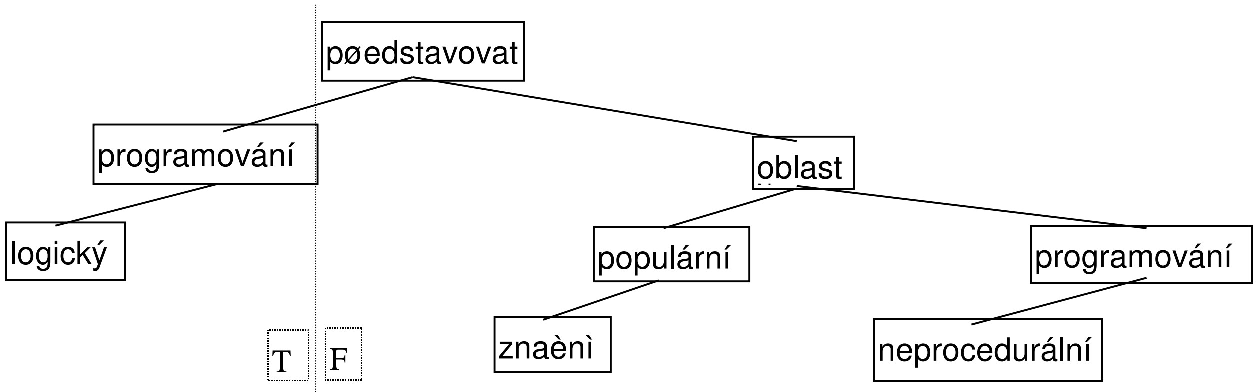
3. TFA-based approach

Our approach is anchored in the Praguian tradition of Functional generative description (FGD, cf. [Sgall 1967], [Sgall et al. 1986], [Petkevič 1987]). According to this theory, the deep (underlying) representation of the linguistic meaning of a sentence reflects its topic-focus articulation (TFA).

In our terms, a discourse consists of text units, roughly equivalent to simple clauses. Every text unit is articulated into its topic- and focus-part (with the possibility that the topic be empty). In utterances consisting of more then one text unit, each text unit has its TFA[7]. Example 4 illustrates the underlying representation of a simple text unit represented by a sketch of a dependency tree, split into the topics and focus.

Example 4

Logické programování představuje značně populární oblast neprocedurálního programování.
[Logical programming represents a quite favourite area of non-procedural programming.]



The TFA of a text unit relates to the discourse context. For the purpose of the present study we concentrate on one aspect of such relatedness, namely how a text unit U_i relates to other text unit(s) preceding it in a discourse by means of a relation between the topic of U_i and a topic, a focus, a sequence of topics and/or foci, or a (sequences of) text unit(s) in the preceding discourse.

3.1. TFA connections and TFChains

A text unit U_i, for i>1, connects to the text unit U_{i-j}, for 1<j<i by means of a *TFA connection*. Following Weil and Daneš and respecting our remarks above, we distinguish three types of TFA connections :

- **parallel** TFA connection: $T(U_i)$ links to $T(U_{i-j})$
- **progressive** TFA connection: $T(U_i)$ links to $F(U_{i-j})$
- **abridged** TFA connection: $T(U_i)$ links to one (or more) preceding text unit(s) of which it is a summarization

where *linking* corresponds to a *contextual relationship* (see Section 4 below). When j=1, the TFA connection can be characterized as *contact*, i.e., utterance U_i is connected to its immediate predecessor(s). Otherwise, we call it a *distant* TFA connection.

The succession of TFA connections throughout a discourse is what we call a *TFChain*. We assume that a discourse is interpreted as coherent if a TFChain can be identified. A TFChain that consists solely of contact TFA connections can be called *continuous*, otherwise a TFChain is *discontinuous*. Usually, a discourse exhibits a discontinuous TFChain in which, however, continuous sub-TFChains can be identified.

Example 5

7Although there are discussions of how to combine the tfa of the smaller text units into the tfa of the utterance as a whole (cf. [Hajièová et al. 1995]), we treat the tfas within complex text units separately in our current treatment.

1. Logické programování $T_1||F_1$ představuje značně populární oblast neprocedurálního programování.
   [Logical programming represents a quite favourite area of non-procedural programming.]
2. Jeho základní princip je $T_2||F_2$ založen na pojmu Hornových klauzulí
   [Its main principle is based on the notion of Horn clauses]
3. a nejznámější jazyky založené na logickém programování jsou $T_3||F_3$ jazyky PROLOG, DATALOG a TABLOG.
   [and the most famous languages based on logical programming are the languages of PROLOG, DATALOG and T ABLOG.]
4. Strukturální programování $T_4||F_4$
   [Structural programming]
5. (∅ $T_5||F_5$ jedná se o něco jiného než strukturované programování)
   (what is concerned is something else than structured programming)
4. tvoří samostatnou oblast neprocedurálního programování.
   [constitutes an separate area of non-procedural programming.]
6. Je $T_6||F_6$ založeno na popisu problému, který se opírá o strukturu vstupních dat nebo o strukturu výpočtu.
   [It is based on a problem description which relies on the structure of the input data or the structure of the computati on.]
7. Teoretickým základem strukturálního programování jsou $T_7||F_7$ atributové gramatiky.
   [The theoretical basis of structural programming are attribute grammars.]
8. Myšlenka strukturálního programování $T_8||F_8$ není nová.
   [The idea of structural programming is not new.]
9. Již Jacksonova metoda se opírá o $F_9||T_9$ popis problému na základě struktury vstupních dat,
   [Already the Jackson's method relied on a description of a problem based on the structure of the input data]
10. ale gramatický formalismus $T_{10}||F_{10}$ nebyl použit.
    [but a grammar formalism was not employed.]

The following TFA connections can be recognized (more exact characterization of TFA connection ty pe in brackets, cf. Section 4.2):

T_2 -- T_1 parallel contact TFA connection (derivation: property);
T_3 -- T_1 parallel distant TFA connection (derivation);
T_4 -- T_1 parallel distant TFA connection (derivation: siblinghood);
T_5 -- T_4 parallel contact TFA connection (derivation: coreference);
T_6 -- T_4 parallel distant TFA connection (coreference);
T_7 -- T_4 parallel distant TFA connection (derivation: property);
T_8 -- T_4 parallel distant TFA connection (derivation: property);
T_9 -- F_6 proggressive distant TFA connection (coreference: substitution);
T_{10} -- F_7 progressive distant TFA connection (derivation: isa);

Because of the distant TFA connections, the TFChain recognized in this discourse is discontinuous. This does not mean that the discourse is incoherent. What it may mean, however, is that there is mor e going on in the structuring of the discourse than we have captured so far[8].

. TFA connections in complex utterances

A complex utterance corresponds to a complex text unit which can be decomposed into smaller, pos sibly still complex, text units with their own TFA. Let us now reproduce Daneš's classification of co mplex utterances in our terms and comment on it:

1. *simple text units* consist of one T-F nexus
2. *conjoined (paratactic) text units* consist of two (or more) simple text units; the pair of simple tex t units can be conjoined in one of the following ways:
   i. conjoined nexuses: $(T_1\text{-}F_1)$ CONJ $(T_2\text{-}F_2)$
   ii. conjoined topics: $(T_1$ CONJ $T_2)$-F which is equivalent to $(T_1\text{-}F)$ CONJ $(T_2\text{-}F)$
       iff F has a distributive interpretation
   iii. conjoined focuses: T-$(F_1$ CONJ $F_2)$ which is equivalent to $(T\text{-}F_1)$ CONJ $(T\text{-}F_2)$
        iff T has a distributive interpretation
3. *condensed (hypotactic) text units* consist of two (or more) simple text units; the pair of simple te xt units can be condensed in one of the following ways:

[8]We have so far conceived of the discourse structure as flat, without embeddings. It is obvious that we need to concern ourselves also with the hierarchical structure of discourse, building on the approaches of Grosz & Sidner [1986] on the one hand and Hobbs [1979], Mann & Thompson [1987] and Asher [1993] on the other. But we shall not do so in the present paper (cf. [Kruijff et al. 1995] and [Kruijff & Schaake forthcoming]).

i. nexus T_2-F_2 is "incorporated" into the topic of nexus T_1-F_1:
   (T_1 COND (T_2-F_2))-F_1, the condensed nexus being denoted as T*-F;
   if T_2=T_1 or T_2=F_1, T_2 can be elided and F2 can be a part of the complex topic
ii. nexus T_2-F_2 is "incorporated" into the focus of nexus T_1-F_1: T_1-(F_1 COND (T_2-F_2)); the conde nsed nexus being denoted T-F*;
   if T_2=T_1 or T_2=F_1, T_2 can be elided and F_2 can be a part of the complex focus

The example below shows a complex text unit 1 with "embedded" text units 2 and 3.

Example 6

1. První autorovi známou prací, $T_1||F_1$
   The first work known to the author
2. která $T_2||F_2$ se zabývá strukturálním programováním
   which is concerned with structural programming
3. $T_3||F_3$ a opírá se o gramatický formalismus (afixové gramatiky),
   and relies on a grammar formalism (affix grammars),
4. je práce Silvarberga (1978)
   is the work of Silvaberg (1978)

The complex utterance in example 6 could be decomposed as follows:

(T_1 COND T_2-(F_2 CONJ F_3))-F_1 (T_3=T_2, and T_3 is elided)

Text unit 3 is conjoined with text unit 2, yielding a complex text unit which is condensed and integr ated into the topic of text unit 1 by linking T_2 with T_1 by means of a grammatical relationship betwe en a relative pronoun, 'která', and a noun, 'práce'.

TFA connections are defined among the smallest possible text units with TFA which cannot be furthe r decomposed[9]. As a result of the analysis of complex text units we therefore obtain several TFA con nections, a sub-TFChain. The complex text unit in The example below shows a complex text unit 1 with "embedded" text units 2 and 3.

Example 6 involves the following TFA connections constituting a continuous TFChain:

T_2 -- T_1 progressive (reiteration); contact TFA connection
T_3 -- T_2 parallel (reiteration realized as deletion); contact TFA connection

We shall close this section by providing a sketch of a TFChain grammar.

3.2. TFChain grammar skeleton

Our TFChain grammar can be seen as a specialized discourse grammar, depicting the perspective of TFA connections throughout a (coherent) discourse. It is as yet severely under-specified, as we have to admit. We present only its context-free backbone here, but obviously, conditions for applying its rules need to be specified, for instance in the spirit of Unification Grammars (cf. [Gazdar & Mellish 1989], [Shieber 1986]).

Context-free Grammar G=(N,T,P,S), where N = {BEGIN, END, TFChain, CONN, PARAL, PROGR,ABBRV,N,‖,T,F} is the set of non-terminals T = {[topic],[focus],[tub],$,#} is the set of terminals S = TFChain is the starting symbol and P is a set of rules defined as follows:	
TFChain --> TFChain CONN N ǀ BEGIN N ǀ TFChain BOUND END	Rule (i)
CONN --> PARAL ǀ PROGR ǀ ABBRV	Rule (ii)
N --> T ‖ F	Rule (iii)
T --> [topic]	Rule (iv)
F --> [focus]	Rule (v) [10]

[9]We are aware that an ambiguity concerning the level at which coordination applies may arise.
[10]The TFChain grammar skeleton G corresponds to TFA connections being defined among simple text units only. If one decides to treat (certain kinds of) complex text units as wholes, one will have to allow for complex topic, T*, and complex focus, F*, and thus modify rules (iii) through (v) along the following lines:

PARAL --> BOUND	Rule (vi)
PROGR --> BOUND	Rule (vii)
ABBRV --> BOUND	Rule (vii)
BOUND --> [tub]	Rule (viii)
BEGIN --> #	Rule (ix)
END--> $	Rule (x)

where [topic] and [focus] stand for the actual topic- and focus-part of the processed text unit, respectively, [tub] stands for a text unit boundary, which may but need not be explicit in the discourse (e.g., comma, periods or other clause boundary markers), and the symbols "$" and "#" are the delimiters of the beginning and the end of the input string, respectively.

Conditions on TFA connections in terms of underlying contextual relationships need to be encoded in the conditions of rule (ii). The application of this rule is responsible for finding an appropriate text unit in the preceding discourse to which the currently processed text unit needs to be linked.

Unfortunately we had to skip an example of a parse-tree illustrating the application of the rules of the grammar G because of lack of space. We shall turn to the relationships underlying TFA connections now.

4. Contextual relationships underlying TFChains

We call the links between expressions and/or the entities they refer to, on which the TFA connections among text units hinge, *contextual relationships*[11]. Contextual relationships can be of many various kinds. Some of them were illustrated in the above examples. We shall now point out contextual relationships at three different levels of representation of a text unit.

4.1. Levels of representation

Contextual relationships are found at all levels of representation, and it is often the case that features represented at different levels contribute to the establishing of a particular contextual relationship between two text units.

The three levels of representation of a text-unit at which we shall discern contextual relationships in the present paper are the representations of the External Form (EF), the Linguistic Meaning (LM) and the Referential Content (RC).

The **External Form** (EF) is the "input form" of a text unit. It corresponds to what is called the graphemic stratum in FGD[12]. The EF-elements are morphemic items, words and expressions. We are concerned with their outer shape.

The **Linguistic Meaning** (LM) of a text unit corresponds to the tectogrammatic stratum in FGD. The LM-elements are thus dependency trees which reflect also the Topic-Focus Articulation (TFA).

The **Referential Content** (RC) of a text unit is a reconstruction of the entities referred to in the utterance, and thus it extends beyond the strata of FGD. The RC-elements are representations of concepts derived from domain knowledge and their relations. They reflect the actual information about the entities conveyed in the text rather than the *general* information available in the DK.

Grammar G':

N --> T*-F*	Rule (iii')
T* --> topic | (T* COND N)-F* | T* CONJ T*	Rule (iv')
F* --> focus | T*-(F* COND N) | F* CONJ F*	Rule (v')

In this approach, some COND's and CONJ's would be treated as TFA connections. However, we shall not tackle these issues currently.

[11] In the approach of Daneš, various types of isotopic (semantic) relationships underlie thematic sequences. Our approach is similar.

[12] The graphemic stratum is considered an alternative to the first out of five strata in FGD, being the phonetic, morpho-phonemic, morphemic, surface syntax, tectogrammatic (cf. [Petkeviè 1987]).

4.2. Types of contextual relationships

We distinguish the following contextual relationships among elements at the individual levels of rep resentation[13]:

. EF-level

The contextual relationships observed among the EFs have the character of relationships between st rings. The EF-elements considered are syn-semantic expressions E_1 and E_2 occurring in text units T_1 and T_2, respectively.

- **Repetition**: E_2 is a literal repetition or a differently inflected form of the same stem as E_1, e.g., ’ modulů’ (noun, gen., pl.), ’moduly’ (noun, nom., pl.), ’modulu’ (noun, gen., sing.)
- **Derivation**: E_2 is a (synthetic or analytic) lexical derivation E_1, e.g., ’gramatický’ (adjective, no m., sing.), ’gramatiky’ (noun, gen., pl.)

. LM-level

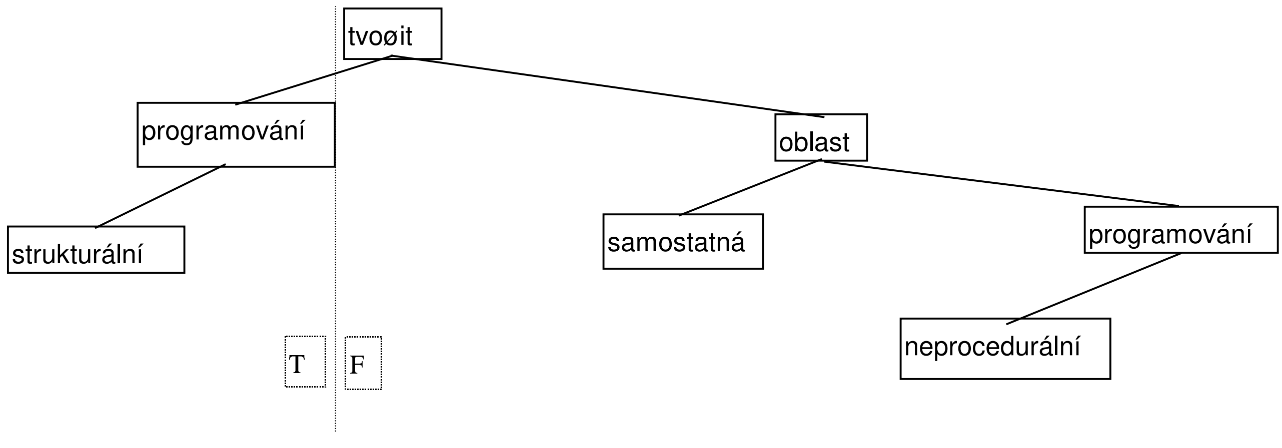
The contextual relationships observed among the LMs have the character of relationships between u nderlying structures of text units represented by dependency trees. Let the dependency tree T_2 corres pond to text unit U_2 and the dependency tree T_1 correspond to text unit U_2:

- **Parallelism**: T_2 parallels T_1 iff they have the same backbone, i.e., there is a homomorphic mapp ing between them preserving the structure of their edges and their labelling. Hence, the structure of T_2 repeats the structure of T_1.
- **Contrast**: T_2 contrasts T_1 iff the best mapping between T_1 and T_2 is one that does not preserve t he structure, but permutes the edges while their labels are the same.

Example 7 *(parallelism)*

1. Logické programování představuje značně populární oblast neprocedurálního programování.
   Logical programming represents a quite favourite area of non-procedural programming.
2. Strukturální programování tvoří samostatnou oblast neprocedurálního programování.
   Structural programming constitutes a separate area of non-procedural programming.

A sketch of the LM-representation of text unit 1 has been shown in Example 4, that of text unit 2 is shown below:



. RC-level

The contextual relationships observed among the RCs have the character of relationships between c oncepts denoted within a text unit.

- **Identity** (ID): identity relation between concepts, e.g., ’menu’ - ’nabídka’

---

[13]We are aware that the specifications of the contextual relationships provided in this paper are sketchy. This is because their formal definitions have not been well-worked out yet.. We beg the readers a pardon and rely on their tolerance.

- **ISA-hierarchy** (ISA): hierarchical relation between denoted concepts, e.g., ’program’ - ’softvér’, ’logické programování’ - ’neprocedurální programování’
- **Contiguity** (C): other-than-hierarchical relation between denoted concepts (e.g., part-of, property-of, etc.), e.g., ’program’ - ’modul’ - ’funkce’, ’modul’ - ’moduly’

4.3. Contextual relationships and TFA connections

In general, as soon as there is a contextual relationship, there is a TFA connection[14]:

contextual relationship	*TFA connection*
EF-repetition(T_k,T_i) or EF-derivation(T_k,T_i)	parallel
LM-parallel(U_k,U_i) or LM-contrast(U_k,U_i)	parallel
RC-identity(T_k,T_i) or RC-isa(T_k,T_i) or RC-contiguity(T_k,T_i)	parallel
EF-repetition(T_k,F_i) or EF-derivation(T_k,F_i)	progressive
RC-identity(T_k,F_i) or RC-isa(T_k,F_i) or RC-contiguity(T_k,F_i)	progressive
EF-derivation(T_k,$U_{i...j}$)	abridged
RC-identity(T_k, $U_{i...j}$) or RC-isa(T_k, $U_{i...j}$) or RC-contiguity(T_k, $U_{i...j}$)	abridged

Essentially, any TFA connection can be either *coreferential* or *non-coreferential*. We distinguish referential relationships as instantiations of conceptual relationships with respect to coreference and non-coreference as follows:

1. Referential relationships with co-reference:
   i. **Reiteration**: two coreferring expressions denote the same concept.
   ii. **Substitution**: two coreferring expressions denote distinct concepts; there is an ISA-hierarchy or a contiguity relation between these concepts.
2. Referential relationships with non-coreference:
   i. **Siblinghood**: two non-coreferring expressions denote the same concept; their referents can be seen as siblings[15].
   ii. **Subordination**: two non-coreferring expressions denote different concepts; there is an ISA-hierarchy relation between the concepts; the referents are distinct, and one of them subsumes the other.
   iii. **Collocation**: two non-coreferring expressions denote different concepts; there is a contiguity relation between the concepts.

Resolution of referential relationships is the core of building an appropriate discourse model, but that is not the subject of this paper, so we shall not go into more detail.[16]

5. Identification of TFChains during text analysis

The identification of a TFChain continuation involves the following steps for every text unit that is to be linked to its preceding context:

- selection of a particular TFA connection-type and formulating an expectation of the referential relations underlying it
- determination of referential relations signified by the contextual relationships of the text-unit at hand by observing the text-unit representation with respect to preceding text-units
- matching the expected referential relations against the signified ones; a positive result means that the TF chain is identified

[14]However, our present formulation of contextual relationships is so all-encompassing that there will always be a contextual relationship between two elements! This is especially due to the conceptual relations of isa and contiguity, but also due to the LM-level relations, unless we are more restrictive about them. More refined specifications of contextual relationships are needed in order to avoid “everything being related to everything”. These issues, however, cannot be further discussed in the present paper.

[15]Compare the notion of “substitution” in [Halliday & Hasan 1976] and that of “identity of sense reference” in [Hirst 1981].

[16]We can say that a hypothesis concerning a TFA connection establishes certain expectations concerning referential relationships. Along with the linguistic clues (like determiners and quantifiers of noun groups) and also taking salience into account, one should have enough ground for resolution. So, we first resolve the rough TFA connection, then we try to resolve the more fine-grained referential relationships, that is, coreference versus non-coreference.

Let us conclude with a sketch on our algorithm for discourse interpretation:

Given: text in a specific domain, domain knowledge-base, the modules for creating an LM from giv en EF (L)[17] and subsequent indication of the concepts denoted in LM (C).

Taking the text utterance by utterance, one text unit at a time, incrementally and sequentially, constr uct the representation at the level of RC, using the TFChain grammar to control the processing, as f ollows[18]:

1. Process the first text unit:
   i. Create a dependency tree using the assumed module L.
   ii. Observe which concepts are denoted by the LM, using the assumed module C.
2. Process every subsequent text unit:
   i. Create a dependency tree using the assumed module L.
   ii. Observe which concepts are denoted by the LM, using the assumed module C.
   iii. Make a hypothesis of TFA connection linking the current text unit to a text unit in the preceding context, using t he TFChain grammar.
   iv. Test the hypothesis by appropriate examining of
      - the EF for contextual relationships to elements in EFs the preceding context
      - the LM for contextual relationships to preceding LM(s)
      - the RC for contextual relationships to preceding RC elements
   v. If the hypothesis has been verified, proceed to step 3.,
      otherwise return to step 2.iii to make another hypothesis.
3. Resolve referential relations:
   i. Determine which referential relations are possible according to the TFA connection at hand.
   ii. Verify whether the referential relations indeed obtain at RC.
   iii. If the verification succeeds, that is, the application of a TFChain grammar rule has been completed and the text unit has been connected to the preceding context, proceed to step 5.
      otherwise, return to step 3.ii to try another possibility for referential relations,
      otherwise return to step 2.iii to make another hypothesis.
4. If no more possibilities for TFA connection are available, interrupt the analysis and allow for intervention.
5. If more text units are to be processed, continue by step 2.
   otherwise quit; the result is a representation of the text-units' contents at the level of RC, integrated by means of TF A connections forming a TFChain.

Due to lack of space, we have to skip an example of the algorithm application.

6. Concluding remarks and further research

7. References

Asher Nicholas [1993]. Referring to abstract objects in discourse. Kluwer Academic Publishers.

Daneš František [1979]. O identifikaci známé (kontextově zapojené) informace v textu [On the identification of known - contextually dependent - information in text]. In: Slovo a slovesnost, Vol. 40: pp 257-270, Prague.

Daneš František [1985]. Věta a text [A sentence and a text]. Studie a práce lingvistické, Vol. 21., Prague.

Gazdar, G. and Mellish, C.S. [1989]. Natural language processing in Prolog: An introduction to computational linguisti cs.

Grosz Barbara, and Sidner Candy L. [1986]. Attention, intentions and the structure of discourse. In: Computational ling uistics, Vol. 12: pp. 175-204.

Hajičová Eva and Sgall Petr [1984]. From topic and focus of a sentence to linking in a text. in Computational models of natural language processing, Guida, G., Bara, B.G. (eds.), Fundamental Studies in Computer Science, Volume 9, Else vier North Holland.

Hajičová Eva, Hoskovec Tomáš and Sgall Petr [1995]. Discourse modelling based on hierarchy of salience. In: Prague Bulletin of Mathematical Linguistics: 64, pp. 5-24. .

[17] Cf. [Kirschner 1987], [Oliva 1989] for systems with similar modules and [Hajièová et al. 1995] for a TFA identification procedure.
[18] Any known parsing algorithm can be in fact applied to do the job, including parallel ones (cf. for instance [Sikkel 1993].

Hajičová Eva, Sgall Petr and Skoumalová Hana [1995]. An automatic procedure for topic-focus identification. In: Com putational linguistics, Vol. 21, No. 1, pp. 81-94.

Halliday Michael A.K. and Hasan R. [1976]. Cohesion in English. London: Longman.

Hirst Graem [1981]. Anaphora in natural language understanding. Springer-Verlag.

Hobbs Jerry R. [1979]. Coherence and coreference. In: Cognitive Science 3(1), pp. 67-90.

Kirschner Zdeněk [1987]. APAC3-2: An English-to-Czech machine translation system. Explizite Beschreibung der Spr ache und Automatische Textbearbeitung, Vol. 13. Prague: Charles University.

Kruijff Geert-Jan and Schaake Jan [forthcoming]. Discerning relevant information in discourses using tfa. In: Recent ad vances in natural language processing, Mitkov R. and Nicolov N. (eds.), John Benjamins Publishers, Amsterdam.

Kruijff Geert-Jan, Schaake Jan and Korbayová Ivana [1995]. Context-independent analysis of multi-sentential turns in i nformation dialogues. Submitted to the International workshop in parsing technologies, September 1995, Prague, The Czech Republic.

Mann William C. and Thompson Sandra A. [1987]. Rhetorical Structure Theory: A theory of text organization. ISI Rep rint Series, ISI/RS-87-190, USC Information Sciences Institute, University of Southern Carolina, Carolina, USA .

Oliva Karel [1989]. A parser for Czech implemented in Systems Q. Explizite Beschreibung der Sprache und Automatis che Textbearbeitung, Vol. 16. Prague: Charles University.

Petkevič Vladimír [1987]. A new dependency based specification of underlying representations of sentences. In: Theor etical linguistics, Vol. 14: pp. 143-172. Berline, New-York: Wlater de Gruyter.

Sgall Petr [1967]. Functional sentence perspective in a generative description. Prague studies in mathematical linguistic s, Vol. 2: pp. 203-225. Prague.

Sgall Petr, Hajičová Eva and Panevová Jarmila [1986]. The meaning of the sentence in its pragmatic aspects. Mey, J.L. (ed.), Reidel.

Shieber Stuart M. [1986]. An introduction to unification-based approaches to grammar. CSLI, Stanford University, Ven tura Hall, CSLI Lecture Notes Number 4

Sikkel Nicolaas [193]. Parsing schemata. PhD Disertation, Twente University, Enschede, The Netherlands.

Weil Henri [1887]. The Order of Words in the Ancient Languages compared with that of the Modern Languages. Trans lated by Super, Charles W. Boston: Ginn & Company, Publishers. New edition with an introduction by Aldo Scaglion e. John Benjamins b.v. (1978). Amsterdam, The Netherlands.