A Corpus-Based Study of Demonstratives in German, Russian and English

Olga Krasavina[1] and Christian Chiarcos[2]

Abstract

The current article presents results from three quantitative corpus studies on the use of demonstrative expressions (demonstrative NPs, demonstrative pronouns) in German, English and Russian. It focuses on two prevalent hypotheses: 1) demonstratives correspond to the medium activation level; 2) demonstratives establish discourse topics. As for (1), it has been repeatedly claimed that referential choice and the activation level of a referent in the memory of the speaker/listener are interconnected (e.g. Chafe 1994). We show that activation is a necessary but not sufficient condition for demonstratives to be used. Strikingly, hypothesis (2), the topic establishment as a function of demonstratives has not received sufficient support from our data for all three languages investigated. As a consequence, we suggest that an appropriate description of the discourse function of demonstratives requires a more detailed study of exemplary contexts beyond the scope of general discourse functions such as topic establishment and givenness marking. Thus, a qualitative study is performed based on the taxonomy of demonstrative functions as suggested by Krasavina (2004). Our results support the assumption that demonstratives prototypically serve as generalized shift-markers in discourse, and thus, both their givenness characterization and their potential to establish new discourse topics can be regarded as epi-phenomenal. This study presents preliminary research for corpus-based studies on multi-lingualism.

1. Background

One of the central isues in the study of discourse anaphora is concerned with the problem of anaphoric distribution in discourse. For any entity to which reference is to be established there is a set of possible anaphoric expressions, all of them equially grammatical. In actual situation of use it is not the case that any of these possible forms is right, or appropriate. Here, the analysts are basically confronted with two perspecives: production and comprehension. In this work, we use the term “referential choice” which reveals the production perspective. In other words, we are concerned with the question: what contributes to the speaker’s choice of an appropriate anaphoric form?

Demonstratives have an important place in the model of referential choice – they are among the most frequently used words in English, Russian and German. By investigating data from these languages, we expect to identify language-independent constraints underlying the choice of demonstratives as compared to pronouns and definite NPs.

[1] Humboldt-University of Berlin
e-mail: krasavio@rz.hu-berlin.de
[2] University of Potsdam, Department of Applied Computational Linguistics
e-mail: chiarcos@uni-potsdam.de

The term “demonstrative” is used to refer to nominal and pronominal expressions with the following structure:

(1) demonstrative NP (DEMNP):a demonstrative determiner, English *this, that*, German *dieser, jener*, Russian *ètot, tot,* followed by a head noun with or without attributes, or

(2) demonstrative pronoun (DEMPRON): a demonstrative, English *this, that*, German *der, dieser, jener*, Russian *ètot, tot*, in nominal use with or without attributes. Note that, here, we abstract from differences between proximal and distal demonstratives, in order to provide a generalized analysis of the discourse function of demonstratives.

Within the approaches of Ariel (1990), Gundel *et al.* (1993), it is claimed that the choice of a referential expression is constrained by the speaker’s evaluation of the referent’s representation in the memory, or activation, in the mind of the listener. Demonstratives express medium activation of a referent, being situated between pronouns (high activation) and definite NPs (low activation). We will refer to this approach as the *mid-activation hypothesis*.

It has also been observed that demonstratives tend to code non-topical referents in German and Russian (*cf.* Diessel, 1999, Bosch *et al.*, 2003). Whereas pronouns code topic continuations, demonstratives refer to referents which are less predictable in the current context. Thus the demonstrative is assumed to support establish references to referents which would not be accessible otherwise. We refer to this approach as the *identificational hypothesis*.

Further, it has been suggested that demonstrative pronouns tend to establish new discourse topics (Schiehlen, 2004). According to Himmelmann (1996), demonstratives are used after the first mention of a referent in order to mark it as “thematically prominent” in the subsequent discourse. We will refer to these approaches as *topic establishment hypothesis*.

Maes and Noordman (1995) argue that, since coreference can be established on activation basis alone, demonstratives are used to modify the existing representation of a referent in discourse, the *modification hypothesis*.

This paper presents a corpus-based approach to clarify to what degree any of these four hypotheses can be held up for the data under investigation. In section 2, a quantitative study on corpus data from German, English and Russian is described which compares the distribution of demonstratives with other referring expressions investigating. In section 3, we present results from a qualititative investigation in which we developed a funcational taxonomy of demonstratives in order to verify the *modification hypothesis* and to identify additional, activation-independent factors determining the use of a demonstrative NP. This taxonomy was applied to instances of DEMNPs in Russian and German. Section 4 presents the Generalized Shift Marking hypothesis and concludes the paper with discussion and results.

2. Quantitative Study

2.1 Corpus Annotation and Feature Extraction

In our analysis, we focused on three corpora of newspaper articles annotated for pronominal and nominal anaphora: 1) 175 texts (33,075 tokens) from the Potsdam Commentary Corpus (Stede 2004), a corpus of German newspaper commentaries

taken from the Märkische Allgemeine Zeitung (MAZ) with very rich annotation; 2) 365 texts (188,035 tokens) from the RST Discourse Treebank (Carlsson *et al.*, 2003), a corpus of English business articles taken from the Wall Street Journal, annotated for discourse structure (Carlsson *et al.*, 2003); 3) 14 Russian articles (45,226 tokens) from a parallel corpus of multilingual news agency articles which is currently compiled by the authors.

The corpora were annotated for co-reference according to the PoCoS core scheme (Chiarcos and Krasavina, 2005) using MMAX[3]. Annotation of the Russian and the German corpus was performed by native-speakers, linguistic students of the University of Potsdam and the Moscow State University.

In the English corpus, we annotated 3412 anaphoric chains covering 11877 referring expressions. In the German corpus, we annotated 864 anaphoric chains covering 2466 markables. And, in the Russian corpus, 106 anaphoric chains covering 416 referring expressions were annotated.[4] In quantitative investigation, we will compare demonstratives with with third-person personal pronouns (shortly "pronoun", PRON), non-demonstrative full NPs (shortly "nominal" or "plain NP", NP), for German and English further divided into definite NP (DEFNP) and indefinite NP (INDNP). Besides these, proper names (NAME) are considered. Other referring expressions have been excluded from our study, thus we considered 9603 referring expressions in the English corpus, 2158 referring expressions in the German corpus, and 380 referring expressions in the Russian corpus.

From the coreference annotation, we extracted the following quantitative properties of demonstratives and control elements within their respective anaphoric chains: 1) **absolute position** within the chain (chain-initial [first mention], chain-final [last mention], chain-medial); 2) **referential distance**, i.e. sequential distance to previous mention; 3) **topic persistence** (frequency of mentions within the next 20 utterances); and 4) **centrality** (length of anaphoric chain relative to the number of utterances of the text).

2.2 Hypotheses

We focus on four hypotheses as to the use of demonstratives: mid-activation, identification, mid-activation, topic establishment and modification. We formulate predictions as to following quantitative characteristics: position in chain (chain-initial, chain-final and chain-medium), referential distance to antecedent (in comparison to plain definite NPs), topic persistence (mentions in the forthcoming discourse), and centrality (frequency of referent mentions), see Table 2. These expectations are elaborated below.

2.2.1 Mid-Activation Hypothesis

Gundel et. al. (1993: 275) argue that each "memory and/or attention state" in the Givenness Hierarchy "is a necessary and sufficient condition for appropriate use of a

[3] http://mmax2.sourceforge.net/

[4] Note that for the Russian corpus, no complete coreference annotation has been undertaken, but a balanced sample of referring expressions was selected in a principled way for which the whole anaphoric chain was annotated.

different form or forms", and that demonstrative NPs correspond to the medium memory and/or attention state (i.e. "activated" or "familiar" statuses).

If this is true, then following predictions must be fulfilled: 1) chain-initial position is dispreferred, chain-final and chain-medial position should be indifferent; 2) for both demonstrative pronouns and demonstrative NPs, distance to antecedent is greater than for pronominal references, for demonstrative NPs, it should be lower as compared to definite NPs; 3,4) topic persistence and centrality should be indifferent.

2.2.2 Identification Hypothesis

If the function of the demonstrative is to clarify reference to otherwise inaccessible referents, we expect 1) insensitivity to chain position; 2) for both demonstrative pronouns and demonstrative NPs, referential distance is greater than for pronouns and definite NPs respectively; 3,4) topic persistence and centrality should be indifferent.

2.2.3 Topic Establishment Hypothesis

According to this hypothesis, demonstratives establish new discourse topics by which we mean discourse referents which a discourse or a discourse segment is *about*, respectively referred to as *global* and *local discourse topic*, *cf.* Lichtenberk (1996).

Thus, the prediction is that demonstrative NPs and demonstrative pronouns are likely to encode referents just introduced into the discourse. Further, this referent has to be mentioned in the further discourse at least once. So, following performance is expected: 1) chain-initial position is not excluded, chain-medial position is preferred, chain-final position is dispreferred; 2) distance for demonstrative pronouns should be higher as compared to personal pronouns, 3,4) topic persistence and centrality of demonstratives should be greater than for non-demonstratives.

2.2.4 Modification Hypothesis

According to Maes and Noordman (1996), a demonstrative NP indicates a modification of an existing representation of a referent, thus it can be predicted that 1) the chain-initial position is dispreferred, chain-final and the chain-medial position should be indifferent; 2) distance should be lower than for defNPs;[5] 3) topic persistence and centrality should be indifferent.

2.3 Results of the Quantitative Study

In Tables 1, 2 and 3 the chain preferences are presented. The results for German and English are, in fact, identical: the distribution of DEFNP and NAME along the anaphoric chain corresponds to the ideal distribution along the anaphoric chain, i.e. they seem to be indifferent to chain position, though proper names show a slight dispreference for the chain-final position. Of course, INDEFNP is restricted to the chain-initial position,

[5] In fact, Maes and Noordman (1995) claim that demonstrative NPs refer to highly activated referents.

which is excluded for pronouns (with the exception of cataphoric pronouns). Rather surprisingly, we find a strong and consistent tendency for DEMPRON and DEMNP to appear chain-finally which is not predicted by any of the hypotheses.

For Russian, the results are similar, with the exception that DEFNP and INDEFNP are not distinguished, thus NP shows the same distributional pattern as DEMNP and INDEFNP in English together. Further, demonstrative pronouns were not considered as there were only 3 instances in the sample.

Based on these observations, we find that INDEFNPs tend to occur chain-initially, PRONs tend to appear chain-medially, and both DEMNP and DEMPRON tend to appear chain-finally, whereas NAMEs and DEFNPs do not seem to show a very concrete preference for a position within a chain.

	DEFNP	DEMPRON	DEMNP	INDNP	NAME	PRON	Total
+ana/-ante	28,18%	5,56%	20,20%	71,05%	39,00%	0,70%	28,73%
+ana/+ante	41,99%	33,33%	28,28%	14,32%	45,37%	50,13%	42,54%
-ana/+ante	29,83%	61,11%	51,52%	14,63%	15,64%	49,17%	28,73%
total	2179	36	99	1292	3703	2294	9603

Table 1 Chain positions per referring expression in English. (Prototypical (> 50%) and atypical (< 15 %) contexts are marked)

	DEFNP	DEMPRON	DEMNP	INDNP	NAME	PRON	Total
+ana/-ante	38,11%	14,58%	23,21%	99,40%	45,09%	1,71%	35,04%
+ana/+ante	28,83%	20,83%	12,50%	0,60%	26,79%	50,57%	29,93%
-ana/+ante	33,06%	64,58%	64,29%	0,00%	28,13%	47,71%	35,04%
	1089	48	56	167	448	350	2158

Table 2 Chain positions per referring expression in German. (Prototypical (> 50%) and atypical (< 15 %) contexts are marked)

	DEMNP	NAME	FULLNP	PRON	Total
+ana/-ante	0,00%	34,65%	51,85%	1,12%	25,48%
+ana/+ante	36,71%	47,52%	37,96%	75,28%	49,04%
-ana/+ante	63,29%	17,82%	10,19%	23,60%	25,48%
	79	101	108	89	377

Table 3 Chain positions per referring expression in Russian. (Prototypical (> 50%) and atypical (< 15 %) contexts are marked)

Consequently, it seems that demonstratives show a behaviour which conflicts with any of the hypotheses above, but especially the topic-establishment hypothesis, as the vast majority of demonstrative descriptions was not mentioned again afterwards, and thus did not establish a discourse topic. These findings are further supported by the measurement of persistence. In English and Russian,[6] DEMNPs and DEMPRON showed the lowest persistence at all, and thus can be regarded as least probable devices to refer to or establish discourse topics, as opposed to plain NPs (general references to concrete objects), NAMEs (often applied to privileged spatio-temporal enities or animate beings) and PRON (preferrably referring to a previously established topic).

[6] The German sample has not been considered in the persistence and centrality study, as the average chain length in the corpus was substantially shorter (2.85) than in the Russian (3.92) and English corpora (3.48), thus limiting the explanatory force of topic persistence.

A related measurement of the anaphoric chain as a whole is centrality, i.e. the number of mentions of a referent relative to the number of utterances. As opposed to persistence, centrality favors referring expressions which refer to the discourse topic rather than establishing a new one. However, both measurements agree in the non-topical characterization of demonstratives, i.e. in the majority of cases they do neither establish new discourse topics (which topic persistence is more sensitive to) nor do they maintain established discourse topics (which centrality is more sensitive to).Since Givón (1983), distance between a referring expression and its antecedent has been recognized as a general measurement of the accessibility of a textually evoked discourse referent. According to the identification hypothesis, the distance between a DEMPRON resp. a DEMNP and its antecedent should be greater than those of PRON resp. NP/DEFNP. However, for DEMNPs, exactly the opposite situation seems to hold, whereas DEMPRONs tend to have a slightly greater referential distance than PRON which is completely consistent with the mid-activation hypothesis.

As a result, we can conclude that none of the four proposals presented above is capable of explaining the quantitative majority of occurrences of demonstrative in the corpora investigated.

The topic-shift hypothesis is contradicted by the end-chain preference of demonstratives, the identification hypothesis predicts a greater referential distance between a DEMNP and its antecedent than for an NP which cannot be observed. The mid-activation hypothesis is compatible with the distance values, though it can neither explain the preference of DEMNPs to refer to non-central referents nor the end-chain preference of demonstratives. In fact, none of the hypotheses does not predict this robust preference of demonstratives to be applied to chain-final referents.

Nevertheless, while it seems that the mid-activation hypothesis is compatible with our distance measurements, it fails to predict the persistence, centrality and chaining effects found in our data, thus, we conclude that givenness in the sense of Gundel *et al.* (1993) possibly specifies *necessary* conditions for the use of demonstratives, but not sufficient ones. Thus, it has to be augmented by another functional dimension. The modification hypothesis of Maes and Noordman (1995) could account for such a functional dimension, though it seems to be incompatible with the end-chain effect and the preference to occur with less central discourse entities as well (at least an additional explanation has to be found why modification should occur among peripheral referents or at the end of an anaphoric chain more often than with discourse topics or in the middle of an anaphoric chain). But as it is defined along qualitative criteria rather than distribution patterns, it cannot be excluded without a deeper investigation of contextual factors and types of modification.

3. Qualitative Study

The results of the quantitative study reported in the last section indicated a clear need for a qualitative analysis of demonstrative uses in our corpora. As a starting point, we derived a taxonomy of demonstrative functions based on classifications provided by Maes and Noordman (1995) and Krasavina (2004), augmented with general assumptions about the general characteristics of demonstratives (*cf.* Diessel, 1999) and hypotheses about the discourse function of demonstrative pronouns (*cf.* Bosch *et al.*, to appear). While classifying our data according to this taxonomy, we introduced

additional distinctions, sharpened definitions and thus, derived an enriched version of this taxonomy in a data-driven fashion. The taxonomy presented in this section is the outcome of this processing and thus to be regarded as a *preliminary* expression of this taxonomy.

In a second processing cycle, the assignment of the previously analyzed data to our taxonomy was verified and corrected. Note that during the classification, we explicitly allowed for examples to be assigned to more than one concept within the taxonomy. While we tried to sharpen out classification, most of these examples could be assigned to just one single category. However, we still do not claim our functional classification to provide *disjoint* categories.

Our dynamic approach on taxonomy development was supported by the application of Protégé,[7] a tool for the development and maintenance of ontologies. The taxonomy was implemented as a conceptual hierarchy, to which examples from the corpus have been integrated as instances. Thus, a knowledge base arose which gives comprehensive overview over all examples previously assigned to one concept in the taxonomy. Accordingly, these examples were taken into account if definitions were adapted or components of the taxonomy were redesigned.

3.1 Functional Taxonomy of Demonstrative Expressions

We will not present a full taxonomy of demonstratives here because of space limitation, a full version is presented elsewhere[8]. For illustration purposes, we will briefly outline the top level categories and selected sub-classes below; the condensed representation of all classes with their respective frequencies is presented in Table 2.

This taxonomy was applied to selected examples from our German and Russian corpora. However, we excluded event anaphors and restricted ourselves to instances of demNP (most pronominal expressions were event anaphors), to consider a more homogeneous data set.

3.2 Application of the Taxonomy to the Corpora

In our investigation of demNPs in German and Russian, we could confirm the observation of Maes and Noordman (1995) that modification is an important type. However, it is important to note that 27,27 percent of the German, and nearly the half (42,24 percent) of the Russian examples cannot be regarded to be modificational because they are based on a plain lexical repetition. Furthermore, one third of the examples in both languages are *trivial* classifications, which means that the anaphor refers to the antecedent by a lexical hypernym, and thus, does not provide any lexical information. Again, it is doubtful whether any “real” modification occurred. We labelled such cases as “trivial classification” in the taxonomy.

[7] http://protege.stanford.edu
[8] http://www.ling.uni-potsdam.de/~chiarcos/

	Frequency in German	Frequency in Russian
Total	44	116
MODIFICATION	32 (72.73%)	67 (57.76%)
ATTRIBUTION	5 (11.36%)	10 (8.62%)
CLASSIFICATION	26 (59.09%)	50 (43.10%)
TRIVIALCLASSIFICATION	14 (31.82%)	40 (34.48%)
PEJORATIVE	4 (9.09%)	6 (5.17%)
TOPICFLOW	8 (18.18%)	43 (37.07%)
INDEFINITETHIS	2 (4.55%)	n/a
REFERENCETOANTITOPIC	1 (2.27%)	16 (13.79%)
ELABORATIONOFLOCALTOPIC	3 (6.82%)	16 (13.79%)
PROMOTETOTOPICSTATE	2 (4.55%)	9 (7.76%)
DEGRADATIONFROMTOPICSTATE	n/a	2 (1.72%)
EXCPLICIT CONTRAST	7 (15.91%)	4 (3.45%)
BINARYOPPOSITION	5 (11.36%)	2 (1.72%)
ONEFROMASET	2 (4.55%)	2 (1.72%)
ATYPICALANAPHOR	8 (18.18%)	2 (1.72%)
VAGUEANTECEDENT	1 (2.27%)	2 (1.72%)
EXOPHORICREFERENCE	6 (13.62%)	n/a
METALINGUISTICANAPHOR	1 (2.27%)	n/a

Table 4: Distribution of functional types of DEMNP anaphors in German and Russian.

MODIFICATION (Maes and Noordman, 1995, Krasavina, 2004)

provide additional lexical material about an already established discourse referent, expression must contain new lexical material, thus non-pronominal, no repetition.

TOPICFLOW (*cf.* mid-activation hypothesis, topic establishment hypothesis)

The demonstrative refers to a non-topical referent or to a referent that changes its topic state, pronominal or lexical repetition.

subclasses: INDEFINITETHIS, PROMOTETOTOPICSTATE , DEGRADATIONFROMTOPICSTATE, REFERENCETOANTITOPIC,ELABORATIONOFLOCALTOPIC

EXPLICITCONTRAST

One referent is contrasted either with another or a class of others, pronominal or lexical repetition.

TEMPORALSHIFT (*cf.* identification hypothesis)

Referent is referred to from a different temporal perspective, pronominal or lexical repetition.

ATYPICALANAPHOR (*cf.* identification hypothesis)

Access to the referent is problematic because it is vaguely defined, not textually evoked or does exist on another meta-level.

subclasses: EXOPHORICANAPHOR, VAGUEANTECEDENT, METALINGUISTICANAPHOR

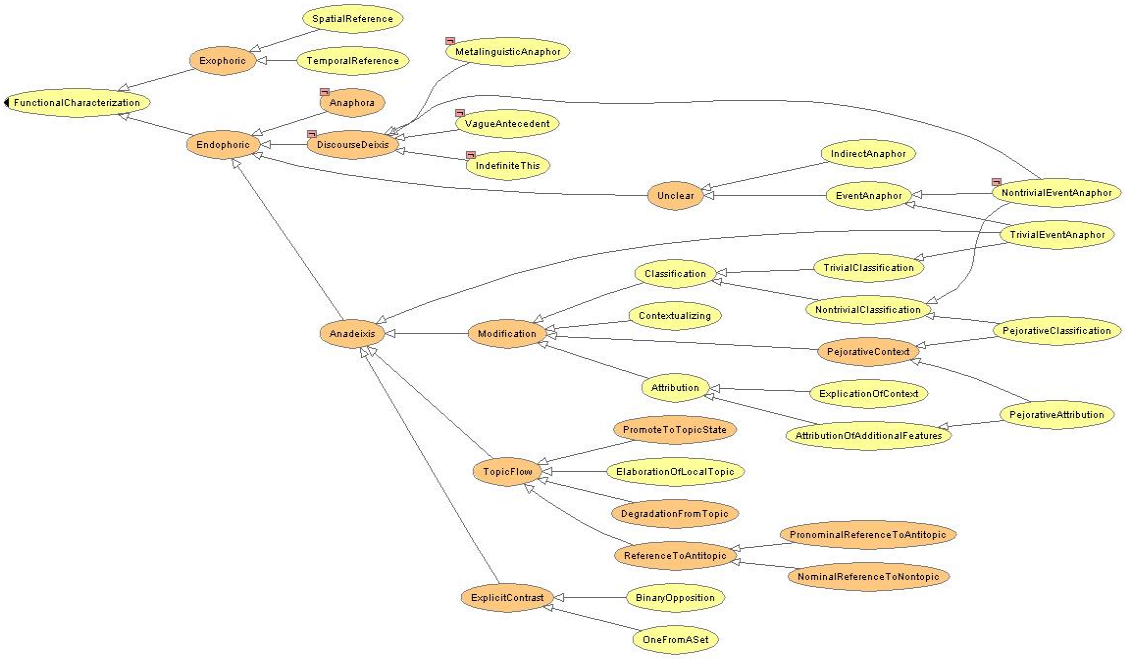

Figure 1: The functional taxonomy of demonstratives

TRIVIALCLASSIFICATION together with non-MODIFICATION make up more than 50 percent of the examples found in the Russian and German corpus, thus indicating that the modification hypothesis cannot be regarded to be the singleton explanation of demonstratives. However, in order to verify whether a 'division of labour' between the modification function and the mid-activation or another alternative exists, we investigated the interplay of the end-chain condition (i.e. chain-final elements) with the taxonomy and the referential distance associated with the functional classes.

3.3 Combining Qualitative and Quantitative Criteria

For this study, we considered 39 German and 65 Russian chance-selected examples from the earlier studies. We found that the end chain condition did not substantially affect the relative frequency of MODIFICATION as opposed to non-MODIFICATION, and thus we confirm our earlier expectation that modification cannot explain the end-chain condition. However, we found that **all** explicit contrast examples in German were chain-final.

In the Russian sample, no instances of explicit contrast occurred. However, we found a tendency for demonstrative NPs to give a final description of a non-topical referent: more than the half of instances of REFERENCETOANTITOPIC appear chain-finally, this type being defined as "picking up a non-salient, but activated element, from a partially ordered set" in our taxonomy. Similar to Bosch *et al.*'s (to appear) analysis of demonstrative pronouns in German, this characterization can be interpreted as based on a discoursally imposed contrasting of the demonstrative with more salient referents. As a first impression, it seems that contrast, either lexically expressed or imposed by the discourse configuration, could explain a certain fraction of the chain-final occurrences of DEMNPs.

Finally, we calculated the average referential distance for any class in our taxonomy, and crucially, we found that MODIFICATION (and each of its sub-classes, including TRIVIALCLASSIFICATION) preferred a greater referential distance than any non-MODIFICATION class (MODIFICATION: German 2.26, Russian 3.12; TOPICFLOW: German 0.4, Russian 0.88; EXPLICITCONTRAST: German: 1.0), the only exception being EXOPHORICREFERENCE (German: 4.0) which can be regarded as distance-insensitive.

Thus, against the prediction by Maes and Noordman (1995), according to which demonstrative NPs are preferably indicate modifications of highly accessible discourse, modification tends to coincide with reference to mid- or low-activated referents.

3.4 A Proposal: Discourse Functions of DEMNPs

As the occurrence of low-distance non-modifications and the end-chain condition are neither explained by the mid-activation hypothesis nor by the modification hypothesis, we must conclude that if applied to highly accessible discourse referents demonstratives have a specialized discourse function beyond their general modification potential.

Besides this, in specialized discourse contexts, demonstratives fulfill certain discourse functions for which we formulate the following hypotheses:

topic promotion

in local contexts, demNPs *establish* or *enforce* the status of newly established discourse topics (INDEFTHIS, PROMOTETOTOPICSTATE, ELABORATIONOFLOCALTOPIC). From our data, this seems to be exclusively licensed to non-central (seldom mentioned) discourse referents.

identification

demonstratives have an *identifying* force and thus can be applied to mid-activated referents (REFERENCETOANTITOPIC), less amenable referents or establish exophoric reference (ATYPICALANAPHOR).

contrast

for chain-final demonstrative NPs, it seems that they are chosen to indicate *explicit contrast* (EXPLICITCONTRAST) or a *discoursally imposed contrast* (REFERENCETOANTITOPIC) between the most salient referent and the referent denoted by the referring expression.

topic demotion

if a demonstrative is used for a referent whose topic status is already sufficiently established, it can indicate a DEGRADATIONOFTOPIC (which is possibly related to the end-chain preference of EXPLICITCONTRAST)

4. Results and Discussion

4.1 Mid-Activation, Modification, and the Functional Classification of Demonstratives

The result of our study are consistent with earlier assumptions that givenness, or more precisely, middle activation, specifies a **necessary** condition for demonstratives (e.g. Gundel et al. 1993, Vieira et al. 2005), but not sufficient ones (Maes and Noordman 1995). For a full account of demonstratives in language, thus, givenness considerations have to be complemented with other functional aspects. Modification seems to be central, but fails to explain the choice of demonstratives for certain discourse configurations.

Based on a functional taxonomy of demonstratives, we suggested four hypotheses for the discourse function of DEMNPs. While topic establishment can be found among these functional types, clear instances of PROMOTETOTOPIC are rare in our corpus, and it seems that instead, demonstrative pronouns and NPs tend to refer to non-central referents which show a lower persistence than those encoded by personal pronouns and definite NPs, and thus serve to establish local discourse topics rather than global discourse topics.

4.2 Towards a Generalization: Demonstratives as Generalized Shift Markers

Beyond this, a general, empirically grounded, explanation of the universal nature of demonstratives has still not come into reach. As a possible candidate, we propose a characterization of demonstratives according to their potential as **generalized shift markers**, or more precisely: demonstratives do not mark the shift itself, but rather underline referent identity whenever a shift occurs. The type of shift can concern:

- shift of reference from the immediate context to the peripheral, cf. *identification*
- shift of perspective, cf. EXPLICITCONTRAST, PEJORATIVE, TEMPORAL SHIFT
- shift of lexical description, cf. MODIFICATION
- shift of local topic, cf. *topic demotion, topic establishment*

This understanding comes close to earlier theoretical models of the licensing conditions of deictic expressions in language. As such, it resembles Ehlich's (1982) conception of **anadeixis**. Based upon a reconstruction of Bühler's (1934) original use

of ``anaphora'' (repeated reference to a previously mentioned referent) and ``deixis'' (identifying an object by explication of its tempospatial location relative to the origo), Ehlich suggests that the text-specific use of demonstratives, i.e. those with textual antecedents, is essentially a form of deixis, that, however, operates on shared knowledge, a common focusing space, rather than on a deictic space. Whereas expressions which are inherently anaphoric (definite descriptions, pronouns) can be applied to a textually evoked entity by retaining its current degree of attention, anadeictic (demonstrative) expressions involve the focusing of the hearer's attention towards this specific entity, i.e. a re-adjustment of attentional states.

However, the band-width of potential discourse functions demonstratives can be ascribed in our corpora suggests that the nature of this "re-adjustment of attentional states" is more complex than only a focusing towards a specific entity. Especially, it seems to involve focusing towards the use of the entity within a specific *context*. Hence, we suggest a broader formulation of the discourse function of demonstratives by arguing that the prototypical function of demonstratives is to indicate object identity under circumstances where other referring expressions would be insufficient. Consistent with Ehlich's anadeixis which involves a transition of deictic functions to the domain of anaphora, this shift-marking potential can be explained ontogenetically by 1) originally deictic properties of demonstratives to pick up an object of a set, 2) their capacity to be used under different activation conditions, and 3) the interaction between pragmatically motivated application of grammatical devices outside their prototypical function (e.g. in order to trigger quality implications, cf. Gundel et al. (1993)) and the potential of grammaticalization to fossilize discourse contexts in which grammatical devices occur (cf. Ariel 2006).

4.3 Other Functions of Demonstratives

The generalized shift marking hypothesis is consistent with earlier proposals, e.g. scenario shift (Sanford and Garrod 1981), rhetorical shift (Fox 1987), episode shift (Tomlin 1987), *inter alia*.

Beyond this, often a pejorative effect of demonstratives is assumed (Kirsner et al. 1987, Krasavina 2004). Here, we'd like to argue that also this pejorative reading can be explained by as a shift in perspective, i.e. from a neutral, objective description towards a highly subjective way of representation. In our account, not the demonstrative itself marks the pejorative character of the description, but it only indicates a perspectual shift between the current and the previous description. If the lexical context does not involve any specific cues about the type of shift (i.e. trivial classification or no modification) which is often the case for pejorative demonstratives, the anaphoric accessibility of the referent does not require a more explicit description, the discourse context is not compatible with a rhetorical or episode shift, and a contrastive reading is also unlikely, then an implication is entailed that this shift must be associated with another level of meaning, such as the social environment. Hence, the pejorative reading can be interpreted as a consequence of the shift-marking potential of demonstratives.

4.4 Open Issues

Finally, we have to point out that several aspects necessitate intensified research. At the moment, it is not fully clear how discourse deixis fits into the taxonomy, which is certainly an important function of demonstratives. Nevertheless, it seems that the shift-marking potential of demonstratives may also play a role here, as discourse dexis, i.e. reference to previously mentioned states or events, can be regarded as "atypical anaphor". In fact, it is argued that discourse deixis differs from anadeixis and anaphora in that the entity which is referred to is constructed upon processing the referring expression. However, the antecedent is nevertheless discourse-old, which is exactly the situation for a referent to be hearer-new and discourse-old at the same time. According to Prince (1981), this situation would be highly unlikely, which however, only indicates the orientation of existing frameworks towards more prototypical anaphoric relationships (in Ehlich's sense). Accordingly, the demonstrative

Beyond the investigation of discourse deixis, we must note that for reasons of data sparsity, we could not differentiate between different types of demonstratives in our analysis. In general, it seemed that both proximal and distal demonstratives behaved similarly, though, for the different languages, the portions between proximal and distal demonstratives were imbalanced. For Russian, as one example, the vast majority of demonstrative NPs were (proximal) *etot*-phrases, whereas most demonstrative pronouns were (distal) *tot*-pronouns. For German and English, the situation was more balanced, but in parts, this might have differenced the observed differences between the languages. Thus, larger corpora for quantitative studies are needed.

Finally, the qualitative studies have to be extended towards the English corpus and demonstrative pronouns, and finally, in order to identify *universal* contexts and properties of demonstratives, it is necessary to investigate the distribution of demonstratives in parallel corpora.

4.5 Methodological Remarks

This work evolved in the course of preparation work for building up a multi-lingual parallel corpus of Russian, English, and German. This language combination has received little attention in corpus studies so far, and actually, no manually analyzed parallel corpus of Russian and German exists at the moment. Our approach is potentially extendable to other languages, and involved testing the methods and tuning up the annotation procedure.

In this article, we attempted to document both our results, and the methodology and tools applied during the research. We see this as a contribution towards the development of Rules of Best Practice (cf. Bird and Simons 2002) which have to be developed by the research community as soon as new technical possibilities become available. However, besides presenting possible solutions, it could be worthful to document problems we faced, too.

A central problem in multilingual studies is the extreme heterogeneity of data and annotation under investigation. During the development of the taxonomy, we have made good experiences with the application of modern ontology-maintenance

software (Protégé) to ensure a consistent terminological backbone shared by the people involved in this study.

With respect to the annotation scheme, we ensured that a cross-linguistically applicable and general annotation scheme (Chiarcos and Krasavina 2005, Krasavina and Chiarcos 2007) was applied to all corpora, and thus, the corpora could be processed with the same software. Further, by means of an ontology of linguistic annotations (Chiarcos 2006), it was possible to develop tag-set neutral scripts to create annotation projects from syntactically analyzed data independently of the underlying tagsets.

By taking these steps, we have defined a protocol which allows the easy processing of additional resources for an analysis as those performed in this study. Syntactically analyzed corpus data is converted into MMAX projects with PoCoS scheme which are manually annotated and evaluated independently of the underlying structure. As all of these processing steps are language-independent, this procedure allows for the integration of additional resources into this analysis process regardless of the underlying annotation scheme and languages as long as the linguistic annotation is handled by the ontology.

Last but not least, a more fine-grained analysis differentiating between proximal and distal demonstratives is necessary to augment our study, which, of course, requires the analysis of greater corpora. Having defined a protocol to integrate novel resources, the possible inclusion of additional resources and beyond this, the extension to other languages, is substantially simplified.

Acknowledgements

We would like to express gratitude to Dmitri Zalmanov for supervising annotation of the Russian and the English corpora. We thank Peter Bosch, Francis Cornish, Konrad Ehlich, Alfons Maes and Carla Umbach and the participants of the Workshop on Anaphoric Uses of Demonstratives at the DGfS 2007 for insightful discussions.

References

Ariel, M. (1990). Accessing Noun-Phrase Antecedents, London: Routledge.

Ariel, M. (2006), Maxi-Grammar vs. Mini-Grammar. Paper presented at Conference on

Intersentential Pronominal Reference in Child and Adult Language, Berlin, December 2006.

Bird, Steven and Simons, Gary (2002): Seven Dimensions of Portability for Language Documentation and Description. In: Language 79, 557–82.

Bosch, P., Rozario T. and Y. Zhao. (2003). Demonstrative pronouns and personal pronouns. German der vs. er. In Proceedings of the EACL 2003, Budapest

Bosch, P., Katz, G. and C. Umbach. (to appear). The Non-Subject Bias of German Demonstrative Pronouns. In Monika Schwarz-Friesel, Manfred Consten, Mareile Knees (eds), Anaphors in Texts.

Botley, S. and McEnery, T. (2001). Demonstratives in English. Journal of English Linguistics, 29 (1).

Bühler, K. (1934), *Sprachtheorie. Die Darstellungsfunktion der Sprache*. Fischer: Stuttgart.

Carlson L., Marcu D. and Okurowski M. E. (2003). *Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory*. In J. van Kuppelvelt and R. Smith (eds). Current Directions in Discourse and Dialogue, Kluwer Academic Publishers.

Chafe, W. (1994). Discourse, Consciousness, and Time. The Flow and Displacement of Conscious Experience in Speaking and Writing. Chicago: University of Chicago Press.

Chiarcos, Ch. and O. Krasavina. (2005). Annotation guidelines. POCOS – Potsdam Coreference Scheme.
http://amor.cms.hu-berlin.de/~krasavio/annorichtlinien.pdf (18.05.2006).

Chiarcos, Ch. (2006), An Ontology for Heterogeneous Data Collections, Proc. Corpus Linguistics 2006, October 10–14, 2006, St.-Petersburg, 373–80.

Cornish, F. (1999). Anaphora, Discourse and Understanding. Evidence from English and French. Oxford: Clarendon.

Diessel, H. (1999). Demonstratives: Form, function and grammaticalization. Typological Studies in Language, 42. Amsterdam: John Benjamins.

Fox, B. A. (1987). Discourse Structure and Anaphora: written and conversational English. Cambridge: CUP

Givón, T. (ed.). (1983). Topic continuity in discourse: an introduction. In Topic continuity in discourse: a quantitative cross-language study, Amsterdam: Benjamins.

Grosz, B., S. Weinstein and A. Joshi (1995). Centering a framework for modeling the local coherence of discourse. Computational linguistics 21(2).

Gundel, J., Hedberg N., Zacharski R. (1993). Cognitive status and the form of referring expressions in discourse. Language 69 (2).

Himmelmann, N. P. (1996) Demonstratives in Narrative Discourse: a taxonomy of universal uses. In B. Fox (ed.), Studies in Anaphora. Amsterdam:Benjamins.

Kibrik, A. A. (1999). Reference and working memory: Cognitive inferences from discourse observation. In K. van Hoek, A. A. Kibrik and A. Noordman (eds.). Discourse Studies in Cognitive Linguistics. Amsterdam: Benjamins.

Kirsner, R.S., V.J. van Heuven, and J.F.M. Vermeulen (1987), Text-type, context and demonstrative choice. *Text* (7): 117–44.

Krasavina, O. (2004). Upotreblenije ukazatel'noj imennoj gruppy v russkom pis'mennom narrativonom diskurse. (The use of demonstratives in Russian narrative discourse). Voprosy jazykoznanija, 3.

Krasavina, O. (2005). Types of third-person pronouns and salience conditions. In Proceedings *Multidisciplinary approaches in discourse: Salience in Discourse* (Stede et al. (eds.)).

Krasavina, O. and Ch. Chiarcos (2007), PoCoS – Potsdam Coreference Scheme. Proc. of the Linguistic Annotation Workshop, held in conjunction with ACL-07. Prague, Czech Republic, June 2007, p. 156–63.

Lichtenberk, F. (1996). Patterns of Anaphora in To'aba'ita Narrative Discourse. In B. Fox (ed.), Studies in Anaphora. Amsterdam:Benjamins, p. 379–411

Maes, A. and L Noordman. (1995). Demonstrative nominal anaphors: a case of nonidentificational markedness, Linguistics 33.

Poesio, M. and N. Modjeska (2005). Focus, activation and this-noun phrases: An empirical study. In A. Branco, T. McEnery and R. Mitkov (eds.), Anaphora

Processing: Linguistic, Cognitive and Computational Modelling., Amsterdam: John Benjamins.

Prince, E.F. (1992), The ZPG letter: subjects, definiteness, and information-status, in S. Thompson and Mann, W.(eds.), *Discourse description. Diverse analyses of a fund raising text*. Amsterdam: John Benjamins.

Sanford, A.J. and Garrod, S.C. (1981). *Understanding Written Language*. Chichester: Wiley.

Schiehlen, M. (2004). Optimizing Algorithms for Pronoun Resolution. In: Proceedings of the 20th International Conference on Computational Linguistics, Geneva, August, 2004.

Stede, M. (2004). The Potsdam Commentary Corpus. ACL-04 Workshop on Discourse Annotation, Barcelona, July.

Tomlin, R.S. (1987). Linguistic reflections of cognitive events. In *Coherence and Grounding in Discourse*, R.S. Tomlin (ed.), 455–80. Amsterdam: Benjamins.

Vieira, R., Salmon-Alt, S. and C. Gasperin (2005). Coreference and anaphoric relations of demonstrative noun phrases in multilingual corpus. In A. Branco, T. McEnery and R. Mitkov (eds.), Anaphora Processing: Linguistic, Cognitive and Computational Modelling. John Benjamins, Amsterdam.