Tatian Corpus of Deviating Examples

T-CODEX 2.1

Corpus Description

Svetlana Petrova
Project B4 in Collaborate Research Center on Information Structure

in cooperation with Carolin Odebrecht, student researcher of project B4

Humboldt University Berlin
## Contents

1 Preliminaries ......................................................................................................................................................3

2 Tagset Declaration ...........................................................................................................................................4

2.1 Layer I: Source data ..................................................................................................................................4

2.2 Layer II: Morpho-syntactic annotation....................................................................................................4

2.2.1 POS – Part of Speech........................................................................................................................5

2.2.2 CAT – Syntactic Category ..............................................................................................................6

2.2.3 CLAUSE STATUS...........................................................................................................................7

2.2.4 GF – Grammatical Function ..........................................................................................................8

2.2.5 Syllable number ...............................................................................................................................9

2.3 Layer III: Information Structure ...........................................................................................................9

2.3.1 GIVENNESS.......................................................................................................................................9

2.3.2 TOP-COMM.......................................................................................................................................10

2.3.3 FOC-BG – Focus-Background .....................................................................................................11

2.3.4 Context...............................................................................................................................................12

2.4 Meta-Information.......................................................................................................................................12

2.4.1 COMMENT .......................................................................................................................................12

2.4.2 BIBLIOGRAPHY ...........................................................................................................................12

References.............................................................................................................................................................13
1 Preliminaries

The present corpus, the Tatian Corpus of Deviating Examples T-CODEX 2.0, provides morpho-syntactic and information structural annotation of parts of the Old High German translation attested in the MS St. Gallen Cod. 56, traditionally called the OHG Tatian, one of the largest prose texts from the classical OHG period. This corpus was designed and annotated by Project B4 of Collaborative Research Center on Information Structure at Humboldt University Berlin.

The OHG Tatian text is a gospel harmony translated from Latin and written down in the middle of the 9th century by at least 6 scribes. In the MS, as Picture 1 shows, the Latin source and the OHG translation are attested as two juxtaposed columns. It has been noticed that each line in the OHG text translates exactly the same material found in the corresponding Latin line (cf. Masser 1997 a and b). A new diplomatic edition made available by Masser (1994, cf. Picture 2) reflects these major characteristics and makes it possible to compare the source and target text. We are grateful to Prof. Masser (Innsbruck) as well as to Vandenhœck & Ruprecht publishing company (Göttingen, Germany) for allowing us to publish parts of the diplomatic text edition by Masser (1994) in our corpus.

A fundamental question concerning medieval translations from Latin is whether the attested word order patterns and constructions are influenced by the syntax of the original or rather representative for the native grammar of the target language. Recent work on the syntactic analysis of OHG has established as a methodological precondition the restriction on clauses that differ from the order of the Latin original (Fleischer 2006). Such instances are called deviating examples. They are generally estimated as evidence for genuine OHG structures (cf. Dittmer and Dittmer 1998, Donhauser 1998). A detailed examination of such examples has shown that the deviations from the Latin occurring in the OHG Tatian part are not random but apply systematically throughout the text (cf. Fleischer, Hinterhölzl and Solf 2008).

The present corpus compiles ca. 2,000 deviating examples found in the text portions of the scribes α, β, γ and ε. Each clause structure represents an extra file annotated with the annotation tool EXMARaLDA and searchable via ANNIS, a general-purpose tool for the publication, visualisation and querying of linguistic data collections, developed by Project D1of the Collaborative Research Center on Information Structure at Potsdam University.
2 Tagset Declaration

The objectives and design principles of a pilot version of T-CODEX 1.0 are described in Petrova et al. (2009).

2.1 Layer I: Source data

In this layer, we provide a line with the Latin source and the OHG translation of the text. In a separate alignment layer, we provide information which OHG item corresponds to which item in the Latin line. Please note that only the OHG text is annotated. In case two words are spelt together in the original, without a space in-between, e.g. because of clitisation, we treat the two words as separate items but place ‘+’ on the clitic, cf. (1):

(1)  
niu+ uuaari  
NEG.PRT-be.3.Sg.PAST.SBJ

2.2 Layer II: Morpho-syntactic annotation

Below we provide a list of the tags used in T-CODEX 2.0. The tag-set declaration is organized alphabetically for each layer of annotation. In general, we aimed at sticking to the canonical definitions of the terms applied. Whenever we apply specific notions or use
common terms in a non-canonical way, we provide additional explanation in the third column.

### 2.2.1 POS – Part of Speech

#### 2.2.1.1 Inflected parts of speech

<table>
<thead>
<tr>
<th>Tag</th>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>adjective</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>noun</td>
<td></td>
</tr>
<tr>
<td>NUM</td>
<td>numeral</td>
<td>Numeral is used as a cover term for cardinal numbers, ordinal numbers and some special classes like 'both', 'lots of', 'half' (cf. Helbig and Buscha, 1987, 320 f.) Eg.: <em>zuelif iaro</em> 'twelve years', T_096_31-32</td>
</tr>
<tr>
<td>PRONDEM</td>
<td>demonstrative pronoun</td>
<td>In OHG, demonstrative pronouns are also used as definite determiners.</td>
</tr>
<tr>
<td>PRONINDEF</td>
<td>indefinite pronoun</td>
<td>This is a cover term for indefinites and quantifiers in OHG including items like 'many', 'all', 'some' etc. This classification reflects the standard interpretation of these classes of words in traditional grammar books on OHG.</td>
</tr>
<tr>
<td>PRONINTERR</td>
<td>interrogative pronoun</td>
<td></td>
</tr>
<tr>
<td>PRONPRS</td>
<td>personal pronoun</td>
<td></td>
</tr>
<tr>
<td>PRONPOS</td>
<td>possessive pronoun</td>
<td>According to the standard view, there is no formal class of possessive pronouns in OHG. We assign this tag to the genitive form of personal pronouns used to express possessive relations.</td>
</tr>
<tr>
<td>PRONRFL</td>
<td>reflexive pronoun</td>
<td></td>
</tr>
<tr>
<td>PRONREL</td>
<td>relative pronoun</td>
<td>In OHG, there is no formal class of relative pronouns. We assign this tag to different demonstrative pronouns used to introduce a relative clause.</td>
</tr>
<tr>
<td>PRONREZ</td>
<td>reciprocal pronoun</td>
<td></td>
</tr>
<tr>
<td>VAUX</td>
<td>auxiliary verb</td>
<td>This tag is assigned to the verbs <em>be</em> and <em>become</em> in combination with the past participle of the main verb when used to translate a Latin passive. The third of the modern German auxiliaries, the verb <em>have</em> as a formal constituent of the perfect construction, is rare in this function in the OHG Tatian.</td>
</tr>
<tr>
<td>VCOP</td>
<td>copula verb</td>
<td>A copula is a verb that links the subject or an object to another phrase, adjective etc. which is part of a complex predicate. The complement of a copula is annotated as PRDNOM at the level of grammatical function (GF).</td>
</tr>
<tr>
<td>Tag</td>
<td>Type</td>
<td>Definition</td>
</tr>
<tr>
<td>-----</td>
<td>------</td>
<td>------------</td>
</tr>
<tr>
<td>VMOD</td>
<td>modal verb</td>
<td>Modals are a class of verbs that are used most often with an infinitive to express ability, disposition or obligation. Pay attention to the fact that the list of modal verbs in OHG is larger than in modern German.</td>
</tr>
<tr>
<td>V</td>
<td>main verb</td>
<td>This tag is assigned to the main lexical verb in the clause.</td>
</tr>
<tr>
<td>VN</td>
<td>verbal noun</td>
<td>According to the standard literature the category of 'verbal noun' in Old High German is a cover term that comprises non-inflected and inflected infinitives (gerund) and the two participles, the present and past participle. (Braune 2004, § 301.5, p.257) E.g.: non-inflected infinitive: ziohan 'to draw, to pull'; inflected infinitive after a preposition (zi) which selects the dative case (gerund): zigebanne, T_031_23; present participle: sehenti 'seeing/ looking', T_026_21; past participle: gitruobet 'aggrieved, saddened', T_026_20</td>
</tr>
</tbody>
</table>

2.2.1.2 Non-inflected parts of speech

<table>
<thead>
<tr>
<th>Tag</th>
<th>Type</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADV</td>
<td>adverb</td>
<td>This tag is assigned to adverbs of all kinds, i.e. to adverbs of time, place, incl. location and direction, and to adverbs of cause and manner.</td>
</tr>
<tr>
<td>CONJ</td>
<td>conjunction</td>
<td>This tag is assigned to conjunctions in general. We do not differentiate at this level between coordinating and subordinating conjunctions.</td>
</tr>
<tr>
<td>INTERJ</td>
<td>interjection</td>
<td>In this category we subsume words that express sentence negation in Old High German, for example the negative particle ni/ ne which cliticizes on the finite verb, like in Ni quam ih 'I didn't come' or words like nalles 'and not', noh 'neither'.</td>
</tr>
<tr>
<td>NEG</td>
<td>negation word</td>
<td>Non-inflected parts of speech with no proper denotation, but special functions which contribute to the truth value of the sentence. E.g: uuârlîhho, ia, enu</td>
</tr>
<tr>
<td>PTC</td>
<td>particle</td>
<td></td>
</tr>
</tbody>
</table>

P | preposition |

2.2.2 CAT – Syntactic Category

<table>
<thead>
<tr>
<th>Tag</th>
<th>Type</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP</td>
<td>adjective phrase</td>
<td></td>
</tr>
<tr>
<td>ADVP</td>
<td>adverbal phrase</td>
<td></td>
</tr>
<tr>
<td>CP</td>
<td>sentential complement</td>
<td></td>
</tr>
<tr>
<td>NP</td>
<td>nominal phrase</td>
<td></td>
</tr>
<tr>
<td>PP</td>
<td>prepositional phrase</td>
<td></td>
</tr>
<tr>
<td>VP</td>
<td>verbal phrase</td>
<td></td>
</tr>
</tbody>
</table>
### 2.2.3 Clause Status

Apart from the traditional classification of main and subordinate clauses, we distinguish causal and relative clauses as separate classes for the reasons described below.

<table>
<thead>
<tr>
<th>CAUSAL</th>
<th>Causally introduced by the causal conjunctions <em>uuanta</em> or <em>bithiu uuanta</em> 'because' in Old High German are ambiguous between main and subordinate clauses. For this reason, they are classified as causal uniformly.</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAINDECL</td>
<td>This tag is assigned to main declarative clauses.</td>
</tr>
<tr>
<td>MAINIMPER</td>
<td>This tag is assigned to main clauses which use the imperative mood to utter requests.</td>
</tr>
<tr>
<td>MAININTERR</td>
<td>This tag is assigned to main interrogative clauses. No further distinction is made between constituent (wh-questions) and propositional questions.</td>
</tr>
<tr>
<td>MAINREQUEST</td>
<td>This tag is assigned to main clauses which use the subjunctive mood to utter requests.</td>
</tr>
<tr>
<td>MAINOPTAT</td>
<td>This tag is assigned to clauses expressing wish and desire which contain the verb in the subjunctive mood.</td>
</tr>
<tr>
<td>REL</td>
<td>This tag is given to attributive relative clauses modifying a nominal argument as well as to free relatives (headless relative clauses) which provide the subject or the object of the matrix clause. Note that attributive relative clauses introduced by demonstrative pronouns in relative function in Old High German are sometimes ambiguous between main and subordinate clauses. Cf. sum tuomo uua In sumero burgi/ ther niforhta got (T 200, 30-31) lat. Iudex quidam erat In quadam ciuitate/ qui deum non timebat 'a certain judge was in a certain town DEM NEG-feared God</td>
</tr>
<tr>
<td>SUB</td>
<td>This category comprises finite complement clauses. Adverbial clauses are further subdivided by type (except of causal clauses which are kept separately).</td>
</tr>
<tr>
<td>SUBCOMPAR</td>
<td>This type of subclause expresses comparison.</td>
</tr>
<tr>
<td>SUBCONC</td>
<td>This type of subclause expresses a concessive relation wrt the matrix clause. In modern German, such clauses are introduced by <em>obwohl</em> etc.</td>
</tr>
<tr>
<td>SUBCOND</td>
<td>This type of subclause expresses conditions.</td>
</tr>
<tr>
<td>SUBCONS</td>
<td>This type of subclause expresses consequences.</td>
</tr>
<tr>
<td>SUBFINAL</td>
<td>This type of subclause expresses intentions.</td>
</tr>
<tr>
<td>SUBLOCAL</td>
<td>This type of subclause expresses embedded locations.</td>
</tr>
<tr>
<td>SUBMODAL</td>
<td>This type of subclause expresses manner.</td>
</tr>
<tr>
<td>SUBTEMP</td>
<td>This type of subclause expresses temporal location or extension.</td>
</tr>
</tbody>
</table>
Participial constructions are used in various functions in Old High German, e.g. to specify the circumstances of the matrix event or to modify a noun in the matrix clause. As they lack a finite verb form they are kept apart from finite subordinate clauses.

**infinitival clause**

This tag is assigned to non-finite complement clauses.

### 2.2.4 GF – Grammatical Function

<table>
<thead>
<tr>
<th>Tag</th>
<th>Definition</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADVCAUS</td>
<td>adverbial of cause</td>
<td>This is an adverbial of cause.</td>
</tr>
<tr>
<td>ADVCOM</td>
<td>comitative adverb</td>
<td>This is an adverbial denoting an accompanying person.</td>
</tr>
<tr>
<td>ADVCOND</td>
<td>conditional adverb</td>
<td>This adverb denotes a condition.</td>
</tr>
<tr>
<td>ADVINSTR</td>
<td>instrumental adverb</td>
<td>This is an adverbial of instrument.</td>
</tr>
<tr>
<td>ADVLOC</td>
<td>adverbial of location</td>
<td>This is an adverbial of location.</td>
</tr>
<tr>
<td>ADVMOD</td>
<td>adverbial of manner</td>
<td>This is an adverbial of manner.</td>
</tr>
<tr>
<td>ADVSENT</td>
<td>sentential adverb</td>
<td>This kind of adverb relates to the entire proposition.</td>
</tr>
<tr>
<td>ADVTEMP</td>
<td>temporal adverb</td>
<td>This is an adverbial of time and temporal duration.</td>
</tr>
<tr>
<td>APPOS</td>
<td>apposition</td>
<td>An apposition is a repetition of an argument expression which provides further information, e.g. <em>zacharias sin fater</em> 'Zacharia, his father'.</td>
</tr>
<tr>
<td>ATTR</td>
<td>attribute</td>
<td>An attribute is a modifier to a noun.</td>
</tr>
<tr>
<td>ATTRPRDNOM</td>
<td>attribute to the predicate noun</td>
<td>This attribute modifies a predicate noun.</td>
</tr>
<tr>
<td>DIR</td>
<td>directional argument</td>
<td>This tag is assigned to the directional arguments of verbs of movement, e.g. <em>in b&amp;hleem in santa sie in b&amp;hleem</em> 'And sent them to Bethlehem', T_040_02</td>
</tr>
<tr>
<td>DO</td>
<td>direct object</td>
<td>This tag is forced by the subject matter of the project B4, namely to investigate verb-object orders in OHG. It is of utmost importance to distinguish between tensed and untensed verb forms and to have a quick access to these categories in the annotation. (cf. Pintzuk 1999).</td>
</tr>
<tr>
<td>GERUND</td>
<td>gerund</td>
<td>This tag is forced by the subject matter of the project B4, namely to investigate verb-object orders in OHG. It is of utmost importance to distinguish between tensed and untensed verb forms and to have a quick access to these categories in the annotation. (cf. Pintzuk 1999).</td>
</tr>
<tr>
<td>INF</td>
<td>infinitive</td>
<td>see above</td>
</tr>
<tr>
<td>IO</td>
<td>indirect object</td>
<td>This tag is forced by the subject matter of the project B4, namely to investigate verb-object orders in OHG. It is of utmost importance to distinguish between tensed and untensed verb forms and to have a quick access to these categories in the annotation. (cf. Pintzuk 1999).</td>
</tr>
<tr>
<td>PARTPF</td>
<td>past participle</td>
<td>This tag is forced by the subject matter of the project B4, namely to investigate verb-object orders in OHG. It is of utmost importance to distinguish between tensed and untensed verb forms and to have a quick access to these categories in the annotation. (cf. Pintzuk 1999).</td>
</tr>
</tbody>
</table>
### PARTPR
**present participle**
- see above

### PO
**prepositional object**
- This is a prepositional phrase which is selected as an argument of a verb or an adjective.
- E.g., *in inan in giloubun in inan thó sine iungiron* 'His disciples started to believe in Him.', T_056_10; *fora gote in siu uuarun rehtiu beidu fora gote* 'They were both righteous before God.' T_026_3

### PRD
**predicative attribute**
- This is a traditional term which expresses secondary predication.
- E.g., *blint in thaz her blint uuvrdi geboran* 'That he was born as a blind man.', T_220_17

### PRDNOM
**predicate noun**
- The predicate noun (germ. *Prädikatsnomen*) is a term used in classical philology to describe the obligatory argument of copula verbs like 'to be', 'to become', 'to call', etc. The predicate noun can be a noun, an adjective, a prepositional phrase and even a whole sentence. E.g., *uuituua* in *thiu uuas uuituua* 'She was a widow.'; *cephas in thu bist giheizzan cephas* 'Your name will be Cephas.'

### SUBJ
**subject**
- The nominative NP which agrees to the finite verb in the clause

### VFIN
**inflected verb**
- This tag at the level of grammatical function is forced by the subject matter of the project B4, namely to investigate verb-object orders in Old High German. It is of utmost importance to distinguish between tensed and untensed verb forms and to have a quick access to the finite verb in the annotation. (cf. Pintzuk 1999)

### VOC
**vocative expression**
- An expression referring to a person to which the utterance is addressed, e.g. *truhtin, meistar* or *fater*.
- The vocative expression typically occurs outside of the clause and not in an argument position selected by the predicate.

#### 2.2.5 Syllable number

Here, we note the number of syllables per syntactic constituent.

#### 2.3 Layer III: Information Structure

##### 2.3.1 GIVENNESS

The following tags are assigned to discourse referents only. For more details on entities that have to be annotated at this level, see the Annotation Guidelines of the SFB, Part on Information Structure, (Dipper et al. 2007, Section 3.2.1.). We assign the following tags:
ACC accessible This tag is assigned to expressions that are accessible via contextual reference to a given entity or to the assumed world knowledge of the interlocutors.

GIV given This tag is assigned to expressions that refer to entities with a given antecedent in the context.

NEW new This tag is assigned to expressions that refer to entities with no antecedent in the context.

2.3.2 TOP-COMM

Annotation at this level applies if the sentence favours an unambiguous topic-comment reading. We decompose the notion of topicality into various features that are related to topics in the literature. Next to informational status annotated separately at the level of GIVENNESS, we especially look at the application of standard topic-tests (2.3.2.1), the use of topic-marking constructions (2.3.2.2), the use of morphological topic markers (2.3.2.3) and the semantic type of NPs (definite vs. indefinite specific vs. existential reading of NPs, cf. 2.3.2.4)

TOP topic If the sentence favours a topic-comment reading as described above, we mark the constituent that is put in A for topic.

COMM comment The rest of a clause with a topic is annotated for comment. If the entire clause is annotated for comment, this implies that a topic-comment division applies but there is no overt topic phrase (e.g., due to topic-drop etc.).

2.3.2.1 ABOUTNESS

This tag is given to phrases that are potential candidates for topics in the clause. With much current work, we distinguish between frame-setting and aboutness topics as instances of an overall notion of topicality, cf. Jacobs (2001).

FRAME frame This tag is assigned to frame-setting topics, i.e. to expressions that reflect identifiable time spans and locations within which the predication holds. For more details, see Annotation Guidelines of the SFB (Dipper et al. 2007), Sec. 4.2.3.

AB aboutness This tag is given to expressions that are possible candidates for A in some well-known topic tests like “As for X, A says that X...” or “ A says about X that X...”.
2.3.2.2 POSITION: TOPIC-MARKING CONSTRUCTIONS

In this tire, we mark the occurrence of a topic-marking construction. We follow the literature in assuming that left dislocation (LD) and hanging topic (HT) are typical topic-marking strategies, though differences in the pragmatic role of both are considered. Additionally, we distinguish the clause-initial constituent of a V2 main clause (German Vorfeld = GV). This distinction only concerns root declaratives. A consistent extension over subordinate clauses will be provided in future.

<table>
<thead>
<tr>
<th>LD</th>
<th>left dislocation</th>
<th>In this construction, a left dislocated element is resumed by a pronoun in the prefield of the main clause. Note that the pronoun and the left dislocated element agree in case, gender and number.</th>
</tr>
</thead>
<tbody>
<tr>
<td>HT</td>
<td>hanging topic</td>
<td>In this construction, a dislocated element and its resuming pronoun do not necessarily agree in case, number and gender.</td>
</tr>
<tr>
<td>GV</td>
<td>German Vorfeld</td>
<td>This tag is assigned to the clause-initial constituent of a V2 main clause.</td>
</tr>
</tbody>
</table>

2.3.2.3 TOPIC MARKERS

Here, we put in formal markers of topicality, e.g. adverbs, particles or modifiers that licence the topical status of a constituent.

2.3.2.4 DEFINITENESS

At this level, we annotate the semantic properties of phrases with respect to the definite/indefinite distinction. In OHG, the system of definite determiners is not fully developed yet. Therefore, we annotate the semantic interpretation of the phrase despite of the presence of a definite or indefinite marker.

<table>
<thead>
<tr>
<th>DEF</th>
<th>definite</th>
</tr>
</thead>
<tbody>
<tr>
<td>INDEF</td>
<td>indefinite</td>
</tr>
</tbody>
</table>

2.3.3 FOC-BG – Focus-Background

On the definition of focus and background, as well as on the different semantic types of focus, see Annotation Guidelines of the SFB (Dipper et al. 2007), Sec. 5. Please note that second occurrence focus is marked by an index cf_1 and cf_2 in the annotation.
CF contrastive focus
NIF new-information focus
NIFCF This combination of tags suggests that new and contrastive focus co-occurs on the same XP.
CF_01 In cases of multiple foci, we assign indices to the foci belonging to different contrastive pairs.
CF_02 see above

2.3.1 FOC-MARKER

In this layer, we annotate focus sensitive operators like additive also, exclusive only, etc., as well as wh-words.

FOPR focus sensitive operator
WH question word

2.3.4 Context

Here, we annotate the relation of the clause with respect to the preceding context, especially in terms of discourse organization. We mainly distinguish between “situation shift” (Situationswechsel) und “anaphoric reference” (“AR”). Situationswechsel involves a change of time, place, speaker/narrator and participants.

2.4 Meta-Information

2.4.1 COMMENT
This is a line for individual remarks and comments of the annotator.

2.4.2 BIBLIOGRAPHY
We provide the respective information according to manuscript page and line, scribe and concordance notes.
References

Braune, Wilhelm

Dipper, Stefanie, Goetze, Michael and Skopeteas, Stavros

Dittmer, Arne and Ernst Dittmer

Donhauser, Karin

Fleischer, Jürg

Fleischer, Jürg, Roland Hinterhölzl and Michael Solf

Helbig, Gerhard and Joachim Buscha

Jacobs, Joachim

Masser, Achim ed.

Masser, Achim

Masser, Achim

Petrova, Svetlana, Solf, Michael, Ritz, Julia, Chiarcos, Christian, Zeldes, Amir

Pintzuk, Susann

Sonderegger, Stefan