CROCO

LINGUISTIC PROPERTIES OF TRANSLATIONS
A CORPUS-BASED INVESTIGATION FOR THE LANGUAGE PAIR ENGLISH-GERMAN

Title **Kodierung von Metainformation**

Author Annette Klinger, Mihaela Vela, Silvia Hansen-Schirra

Deliverable No. 2
Work package 1.3
Status *final version*
Availability *local*
Date *16 Mai 2006*

DFG project STE 840/5-1
http://fr46.uni-saarland.de/croco/

Kodierung von Metainformation

1 Einleitung

Bei der Erstellung des CroCo-Korpus spielen die Meta-Informationen der einzelnen Texte eine ganz besondere Rolle. Die Sub-Korpora wurden nämlich so zusammengestellt, dass sie in Bezug auf die Meta-Informationen eine möglichst große Bandbreite an übersetzungsrelevantem Sprachgebrauch abdecken. Dies wird insbesondere bei der Auswahl der einzelnen Register deutlich. Deshalb wird jeder einzelne Text durch eine kurze Registeranlayse charakterisiert. Zusätzlich spielen natürlich auch Informationen zum Publikationsdatum und -ort, zum Autor, zur Übersetzungsrichtung etc. eine wichtige Rolle, da auch sie für die Sammlung der Sub-Korpora als Basis dienten (siehe Neumann & Hansen-Schirra 2005 sowie Deliverable No. 1 „Corpus Design" für eine genauere Beschreibung der Kriterien zur Korpuserstellung). Aus diesem Grund ist es wichtig, die Meta-Informationen der einzelnen Texte gut zu dokumentieren. Dies macht die Text-Auswahl transparent und für Projekt-externe Nutzer des Korpus nachvollziehbar. Darüber hinaus kann man das Korpus auch nach diesen Informationen filtern und individuelle, zweck-gebundene Sub-Korpora extrahieren (z.B. nach Sprache, Übersetzungsrichtung, Register, etc.).

Durch die Dokumentation der eigenen Annotation bzw. der eigenen Alignierung und deren Nachbearbeitung (siehe Kapitel 2.3 und 2.4) wird auch der Enkodierungsprozess transparenter. Auf diese Weise können Problemfälle und Korrekturen zurückverfolgt werden, was für die Konsistenz des Korpus sowie der Annotations- und Alignment-Schemata wichtig ist. Hierdurch wird die Nachhaltigkeit der Daten innerhalb des CroCo-Projekts und auch für die spätere Freigabe und Nutzung garantiert.

Im Folgenden werden die unterschiedlichen Meta-Informationen der im Korpus enthaltenen Texte, deren Enkodierung im CroCo-Header und das Format des Headers genauer beschrieben (siehe Kapitel 2). In Kapitel 3 wird dann das Annotationswerkzeug „CroCo-Meta", die in CroCo entwickelte Eingabemaske für die Header-Informationen, vorgestellt.

2 CroCo-Header

Das vorliegende Kapitel beschreibt, wie *File Description* (2.1), *Encoding Description* (2.2), *Annotation Description* (2.3) und *Alignment Description* (2.4) im CroCo-Projekt realisiert werden. Diese Kategorien werden in Anlehnung an den Standard der Text Encoding Initiative (Sperberg-McQueen & Burnard 1994) annotiert. Die annotierten Meta-Informationen werden im im multi-layer stand-off XML-Format gespeichert. Im Anhang ist eine Beispiel-Annotation eines CroCo-Headers zu finden.

2.1 File Description

Die *File Description* dient zur Identifizierung der Korpus-Texte (siehe Abbildung1):

```
<fileDesc>
 <filename/>
 <subcorpus/>
 <language/>
 <titleStmt>
  <title/>
  <author/>
 </titleStmt>
 <translation/>
 <publicationStmt>
  <publisher/>
  <date/>
  <distributor/>
  <availability/>
 </publicationStmt>
 <sourceDesc>
  <author/>
  <title/>
 </sourceDesc>
 <registerAnalysis>
  <register/>
  <field>
   <experientialDomain/>
   <goalOrientation/>exposition
  </field>
  <tenor>
   <agentiveRole/>
   <socialRole/>
   <socialDistance/>
  </tenor>
  <mode>
   <languageRole/>
   <channel/>
   <medium/>
  </mode>
 </registerAnalysis>
</fileDesc>
```

Abbildung 1: *File Description*

Hier werden Identifizierungsdaten, wie Dateiname[1], Sub-Korpus (Register und Sprache), Sprache (English oder German), Titel, Autor (zuerst Nachname, dann Vorname), zu jedem Text angegeben. Weiterhin wird bekundet, ob es sich um eine Übersetzung oder einen Original-Text handelt (wobei für Originale der Eintrag leer bleibt, und ansonsten die Abkürzung für die Übersetzungsrichtung eingetragen wird: E-G für Englisch-Deutsch und G-E für Deutsch-Englisch).
Im *Publication Statement* werden Informationen zu Verlag (bei Web-Seiten der Betreiber), Publikationsdatum (bei Webseiten der Tag der letzten Ansicht), Distributor (bei Webseiten die URL) und Verfügbarkeit (lokal oder öffentlich) des Textes gegeben. Natürlich können nur die Informationen eingetragen werden, die auch bekannt sind.
Die *Source Description* ist nur für übersetzte Texte relevant. Hier werden Titel und Autor des Original-Textes der Übersetzung angegeben. Handelt es sich bei der zu annotierenden Datei um ein Original, bleibt die *Source Description* leer.
Um die Texte linguistisch näher zu klassifizieren, wird für jeden Text eine kurze Registeranalyse durchgeführt. Die Ergebnisse werden in der <registerAnalysis> eingetragen. Als Register kommen die folgenden Kürzel in Frage: TOU (Tourismusbroschüren), ESSAY (Wirtschaftsessays), SPEECH

[1] Der Dateiname setzt sich aus dem Kürzel für die Sprache, Original/Übersetzung und Register sowie der Nummer des Textes und der Dateierweiterung zusammen.

(politische Reden), FICTION (fiktive Erzählungen), SHARE (Aktionärsbriefe), WEB (Internetseiten), INSTR (Instruktionstexte), POPSCI (populärwissenschaftliche Texte).
Darüber hinaus gliedern sich die Registerangaben in <field>, <tenor> und <mode>:

- Im *Field* wird in der *Experiential Domain* das Thema und in der *Goal Orientation* die Funktion des Textes angegeben. Hierbei wird zwischen den folgenden Funktionen unterschieden: *exposition* (darstellend), *narration* (erzählend), *argumentation* (argumentativ), *persuasion* (überzeugend), *instruction* (instruktiv).
- *Tenor* spezifiziert *Agentive Role*, *Social Role* und *Social Distance*. *Agentive Role* beschreibt das Verhältnis zwischen Autor und Rezipient. Hier stehen folgende Angaben zur Auswahl: *expert to expert* (Experten-Kommunikation), *expert to layperson* (Experte-zu-Laie-Kommunikation), *layperson to expert* (Laie-zu-Experte-Kommunikation), *layperson to layperson* (Laien-Kommunikation). Die soziale Rolle gibt Aufschluss über die Stellung des Autoren zum Rezipienten. Hier wird zwischen gleich *(equal)* und ungleich *(unequal)* unterschieden. Die soziale Distanz kann mit folgenden Kategorien spezifiziert werden: *casual* (zwanglos, locker), *neutral* (neutral), *formal* (förmlich), *intimate* (vertraulich), *colloquial* (umgangssprachlich), *consultative* (beratend).
- *Mode* trifft nähere Aussagen über *Language Role*, *Channel* und *Medium*. Die Rolle der Sprache wird durch *ancillary* (Sprache ist untergeordnet, ergänzend) und *constitutive* (Sprache steht im Vordergrund) unterschieden. Beim Channel kann der Text graphisch bzw. gedruckt vorliegen (*graphic*) oder in elektronischer Form vorhanden sein (*electronic*). Medium beschreibt, ob der Text in geschriebener (*written*) oder gesprochener (*spoken*) Form vorliegt oder ob er in geschriebener Form vorliegt, allerdings zum Vortragen konzipiert wurde (*written to be spoken*).

Eine detaillierter Beschreibung der Registeranalyse ist in (Ghadessy 1993) zu finden.

2.2 Encoding Description

Die *Encoding Description* dient zur Dokumentation der Korpuserstellung (siehe Abbildung 2).

```
<encodingDesc>
 <projectDesc/>
 <samplingDesc>
  <extent/>
  <size/>
 </samplingDesc>
 <profileDesc>
  <creation/>
  <langUsage/>
 </profileDesc>
</encodingDesc>
```

Abbildung 2: *Encoding Description*

In der *Project Description* wird das im Korpus verwendete Repräsentationsformat eingetragen. Für die Texte des CroCo-Korpus steht hier „modified tei“, da das CroCo-Repräsentationsformat eine Abwandlung bzw. Erweiterung des Standards der Text Encoding Initiative (Sperberg-McQueen & Burnard 1994) darstellt. Dies gilt für alle Texte des CroCo-Korpus.

In der *Sampling Description* wird beschrieben, aus welchen Bestandteilen das Korpus besteht. Hier wird im Element <extent> der Umfang der Datei angegeben, wobei zwischen „sample“ und „full“ unterschieden wird. Das erste wird gewählt, wenn nur ein Textausschnitt vorliegt; dahinter werden, wenn bekannt, zusätzlich die Seitenzahlen angegeben. Das letztere wird verwendet, wenn der Text vollständig ins Korpus aufgenommen wurde. Im Element <size> wird die Textgröße, gemessen an der Anzahl der gezählten Wörter, angegeben.

Die *Profile Description* gibt zum einen das Erstellungsjahr des betreffenden Sub-Korpus an. Dies wird

im Element <creation> kodiert. Wurden die Sub-Korpora im CroCo-Projekt erstellt, wird als Erstellungsjahr das Jahr 2005 oder 2006 eingetragen[2]. Wurde hingegen ein Sub-Korpus übernommen, wird hier das Jahr der eigentlichen Korpuserstellung eingetragen. Neben dem Erstellungsdatum wird in der *Profile Description* auch noch die Sprache jedes Textes im Element <langUsage> angegeben. Hier werden die nationalen Sprachvarianten unter Verwendung der Ländercodes nach ISO 3166 unterschieden. Im CroCo-Projekt wird für das deutsche Sub-Korpus ausschließlich DE eingetragen, da keine anderen nationalen Sprachvarianten gefunden wurden. Im Englischen wird zwischen der britischen (UK) und der amerikanischen (US) Variante unterschieden. Diese Unterscheidung erfolgt basierend auf einer Untersuchung der Unterschiede zwischen beiden Varianten im Vokabular (z.B. „lift" für UK vs. „elevator" für US) und der Rechtschreibung (z.B. „centre" für UK vs. „center" für US oder „organise" für UK vs. „organize" für US). Kann die Sprachvariante eines Textes nicht eindeutig bestimmt werden, bleibt dieses Element leer.

2.3 Annotation Description

Im CroCo-Header wird für jede Annotationsebene eingetragen, wer die Erst-Annotation durchgeführt hat, aber auch wer für Konstitenzchecks und Korrekturen verantwortlich ist (siehe Abbildung 3). Hierfür wird im Element <respStmt>, was für „responsibility statement" steht, im Element <name> der Name des Annotierers und im Element <responser> der Name des Nachbearbeiters eingetragen. Wird eine Annotation automatisch durchgeführt, wird folgendes eingetragen: <name>automatic</name>.

```
<annotationDesc>
 <word>
  <respStmt>
   <name/>
   <responser/>
  </respStmt>
 </word>
 <chunk>
  <respStmt>
   <name/>
   <responser/>
  </respStmt>
 </chunk>
</annotationDesc>
```

Abbildung 3: *Annotation Description*

Es wird zwischen den Annotationsebenen <word> und <chunk> unterschieden. Auf der Wortebene wird hier die morphologische Annotation dokumentiert. Diese wird in der Erstannotation automatisch durchgeführt, wird dann aber korrigiert. Das Part-of-Speech-Tagging wird in der *Annotation Description* nicht berücksichtigt, da es vollautomatisch vonstatten geht und auch nicht korrigiert wird. Auf der Chunk-Ebene werden sowohl Phrasen als auch grammatische Funktionen manuell annotiert. Diese beiden Annotationsarten werden allerdings immer gemeinsam behandelt, insofern wird auch im Header keine Unterscheidung vorgenommen.

2.4 Alignment Description

Ähnlich wie in der *Annotation Description* wird auch für jede Alignmentebene eingetragen, wer die Erst-Alignierung durchgeführt hat, aber auch wer für Konstitenzchecks und Korrekturen im Alignment verantwortlich ist (siehe Abbildung 4). Hierfür wird – genau wie in der *Annotation Description* – im Element <name> der Name des Alignierers und im Element <responser> der Name des

[2] Die Texte der Sub-Korpora, die innerhalb des CroCo-Projekts erstellt wurden, wurden ausschließlich in der ersten Projektphase im Jahr 2005 gesammelt.

Nachbearbeiters eingetragen. Wird ein Alignment automatisch durchgeführt, wird ebenfalls folgendes vermerkt: <name>automatic</name>.

```
<alignmentDesc>
 <word>
  <respStmt>
   <name/>
   <responser/>
  </respStmt>
 </word>
 <chunk>
  <respStmt>
   <name/>
   <responser/>
  </respStmt>
 </chunk>
 <clause>
  <respStmt>
   <name/>
   <responser/>
  </respStmt>
 </clause>
 <sentence>
  <respStmt>
   <name/>
   <responser/>
  </respStmt>
 </sentence>
</annotationDesc>
```

Abbildung 4: *Alignment Description*

Das Alignment gliedert sich in vier Ebenen: <word>, <chunk>, <clause> und <sentence>. Wort- und Satz-Alignment gehen automatisch vonstatten, gefolgt von einer manuellen Nachbearbeitung. Die Alignierung der Einzelsätze und Chunks wird manuell durchgeführt.

3 CroCo-Meta

Zur benutzerfreundlichen Eingabe der Metadaten in eine Header-Datei wurde die graphische Eingabemaske „CroCo-Meta“ entwickelt. Diese GUI ist in Java implementiert. Mit Hilfe dieses Annotationswerkzeugs kann der Annotierer die oben beschriebenen Kapitel Element für Element abarbeiten. Hierfür kann er die notwendige Information in vorgegebene Felder eintragen bzw. aus einer Liste mit festgelegten Antwortmöglichkeiten wählen (siehe Abbildung 5). Sobald die Meta-Informationen für eine Datei vollständig eingetragen wurden, klickt der Annotierer auf „Submit“. Hierbei erstellt das Programm automatisch eine Header-Datei (generiert aus dem angegebenen Namen der Textdatei und der Dateierweiterung .header) und speichert die eingegebenen Informationen im multi-layer stand-off XML-Format (siehe Anhang).

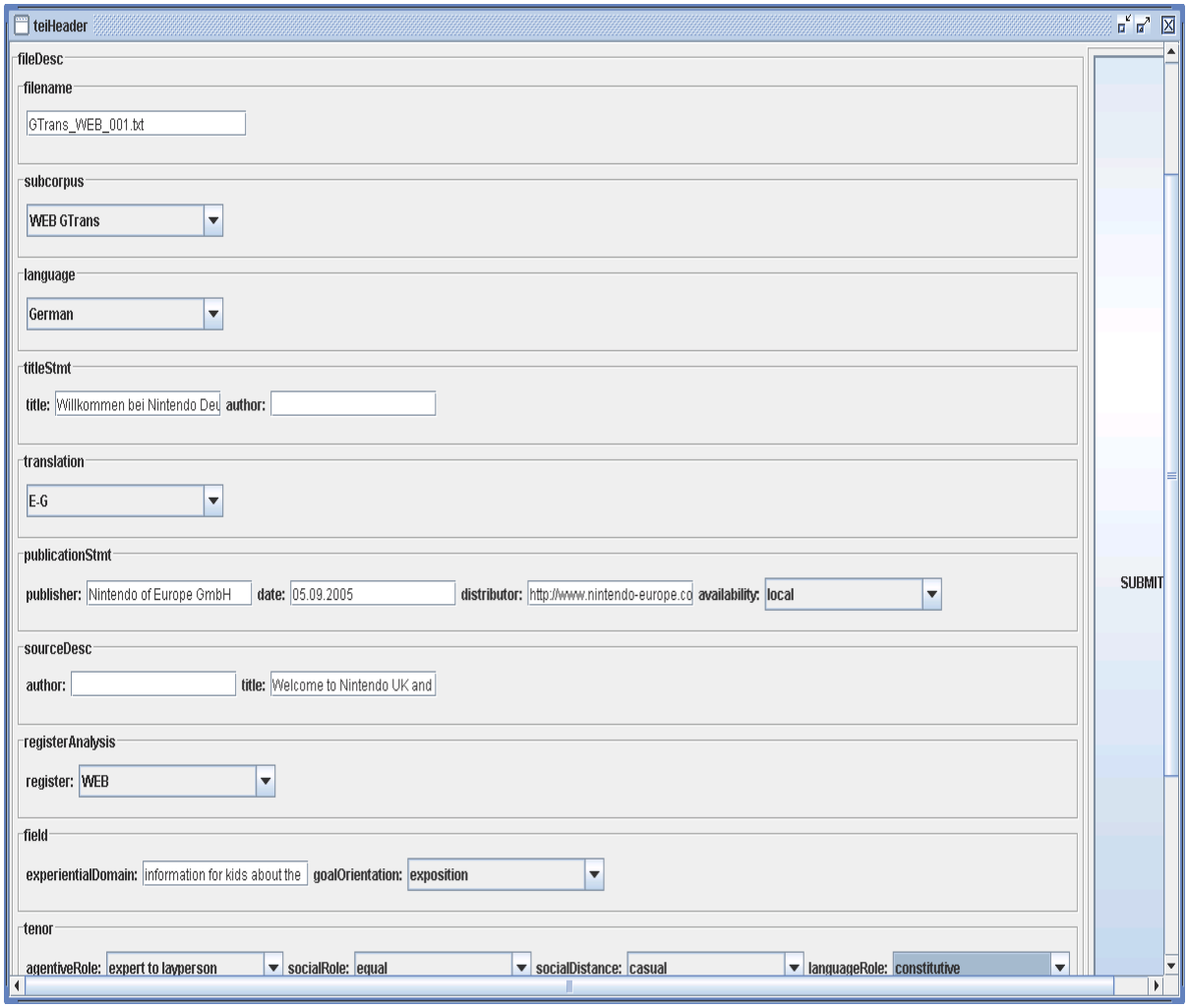

Abbildung 5: CroCo-Meta zur Header-Annotation

Durch die Eingabemaske CroCo-Meta kann auch die Nachhaltigkeit innerhalb des Projekts dokumentiert werden. Abbildung 6 zeigt, dass für jede Alignment- und Annotationsebene festgehalten wird, wer die Erstbearbeitung durchgeführt hat, aber auch wer für Konsistenzchecks und Korrekturen verantwortlich ist (siehe auch Kapitel 2.3 und 2.4). Durch diese Prozessdokumentation werden Problemfälle und Änderungen transparent und lassen sich für spätere Auswertungszwecke zurückverfolgen.

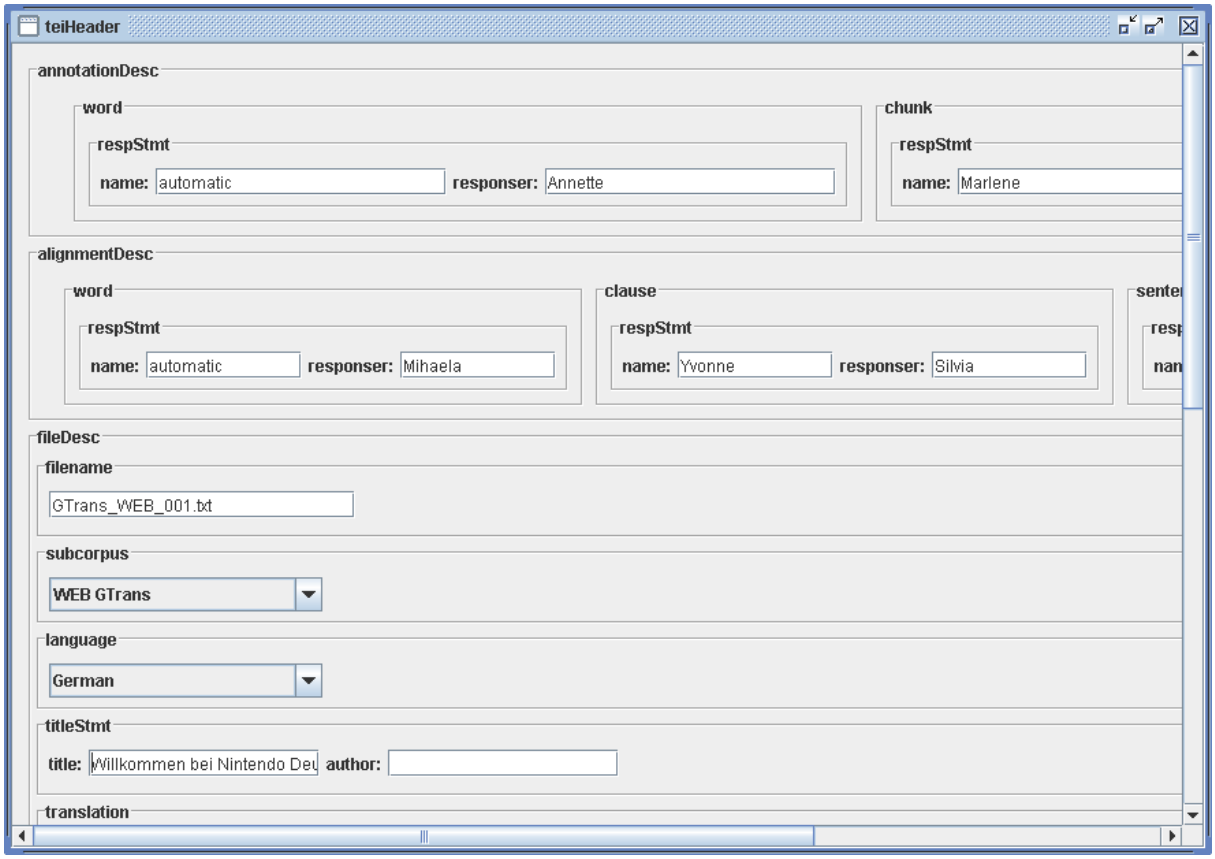

Abbildung 6: CroCo-Meta zur Nachhaltigkeitsdokumentation

Das Werkzeug CroCo-Meta erleichtert die Header-Annotation erheblich, da es dem Annotierer die einzelnen Informationsblöcke als Felder präsentiert. D.h. der Annotierer muss sich nicht in der XML-Datei, die bei zu vielen Elementverschachtelungen schnell unübersichtlich wird, zurechtfinden. Die Vorgabe von Auswahlmöglichkeiten macht den Annotationsprozess schneller und weniger fehlerlastig. Durch die Nutzung von CroCo-Meta wird also die Eingabe der Meta-Informationen benutzerfreundlicher und effizienter. Das Header-Annotationswerkzeug „CroCo-Meta“ ist auf andere Korpusprojekte anpassbar und für wissenschaftliche Zwecke frei verfügbar.

References

Ghadessy, Mohsen, 1993, *Register Analysis, Theory and Practise*. Continuum International Publishing Group.

Neumann, Stella & Silvia Hansen-Schirra, 2005. The CroCo Project. Cross-linguistic corpora for the investigation of explicitation in translations. In: *Proceedings from the Corpus Linguistics Conference Series*. Vol. 1 no. 1, ISSN 1747-9398.

Sperberg-McQueen, Christopher Michael & Lou Burnard (eds.), 1994. *Guidelines for Electronic Text Encoding and Interchange (TEI P3)*. Text Encoding Initiative, Chicago and Oxford. http://www.tei-c.org/

Anhang: Beispiel für CroCo-Header

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE teiHeader SYSTEM "header.dtd">
<teiHeader>
 <fileDesc>
  <filename>GTrans_WEB_001.txt</filename>
  <subcorpus>WEB GTrans</subcorpus>
  <language>German</language>
  <titleStmt>
   <title>Willkommen bei Nintendo Deutschland</title>
   <author/>
  </titleStmt>
  <translation>E-G</translation>
  <publicationStmt>
   <publisher>Nintendo of Europe GmbH </publisher>
   <date>05.09.2005</date>
   <distributor>http://www.nintendo-europe.com</distributor>
   <availability>local</availability>
  </publicationStmt>
  <sourceDesc>
   <author/>
   <title>Welcome to Nintendo UK and Ireland</title>
  </sourceDesc>
  <registerAnalysis>
   <register>WEB</register>
   <field>
    <experientialDomain>information for kids about the privacy policy of
Nintendo, the terms and conditions and the advantages of a
registration</experientialDomain>
    <goalOrientation>exposition, persuasion</goalOrientation>
   </field>
   <tenor>
    <agentiveRole>expert to layperson</agentiveRole>
    <socialRole>equal</socialRole>
    <socialDistance>casual</socialDistance>
   </tenor>
   <mode>
    <languageRole>constitutive</languageRole>
    <channel>electronic</channel>
    <medium>written</medium>
   </mode>
  </registerAnalysis>
 </fileDesc>
 <encodingDesc>
  <projectDesc>modified tei</projectDesc>
  <samplingDesc>
   <extent>sample</extent>
   <size>2.540</size>
  </samplingDesc>
  <profileDesc>
   <creation>2005</creation>
   <langUsage>DE</langUsage>
  </profileDesc>
 </encodingDesc>
 <annotationDesc>
  <word>
   <respStmt>
    <name>automatic<name>
    <responser/>
```

```
   </respStmt>
  </word>
  <chunk>
   <respStmt>
    <name>Marlene<name>
    <responser/>
   </respStmt>
  </chunk>
 </annotationDesc>
 <alignmentDesc>
  <word>
   <respStmt>
    <name>automatic<name>
    <responser/>
   </respStmt>
  </word>
  <chunk>
   <respStmt>
    <name>Marlene<name>
    <responser/>
   </respStmt>
  </chunk>
  <clause>
   <respStmt>
    <name>Yvonne<name>
    <responser/>
   </respStmt>
  </clause>
  <sentence>
   <respStmt>
    <name>Annette<name>
    <responser/>
   </respStmt>
  </sentence>
 </annotationDesc>
</teiHeader>
```