Towards a Cross-Linguistic Production Data Archive: Structure and Exploration*

Michael Götze[1], Stavros Skopeteas[1], Torsten Roloff[1], and Ruben Stoel[2]

[1] SFB 632 "Information Structure", Institut für Linguistik, Universität Potsdam,
Postfach 60 15 53, 14415 Potsdam, Germany
goetze@ling.uni-potsdam.de,
{troloff,skopetea}@rz.uni-potsdam.de,
http://www.sfb632.uni-potsdam.de
[2] Leiden University Centre for Linguistics, van Wijkplaats 4, P.O. Box 9515,
2300 RA Leiden, The Netherlands
R.B.Stoel@let.leidenuniv.nl

Abstract. The present paper presents the structure of a cross-linguistic database of production data. The database contains annotated texts collected from a sample of fifteen different languages by means of identical data gathering methods, which are designed to enable studies on typology and universals of information structure. The special property of this database is that it combines the features of a natural language corpus and the features of a typological database. The challenge for the exploration interface is to provide user-friendly support for exploiting this particular type of resource, thus facilitating empirical generalizations about the collected data in the individual languages and comparison among them.

1 Introduction

The developments of the two past decades have given rise to the creation of a large number of electronic archives of language data. Nowadays there are several typological databases designed to empirically support linguistic comparison across different grammatical systems (e.g., *WALS* in [15] which includes features of all layers of grammar in a large sample of the world's languages, *Autotyp* in [2] which is especially designed to allow for typological and areal generalizations, as well as several databases on particular grammatical domains such as deponency in [6], systems of lexical tones in [16], agreement phenomena in [4], intensifiers and reflexives in [18] and reduplication phenomena in [19]). These resources are designed for the archiving of grammatical features for typological comparisons. Primary data are only available in some of them in the form of illustrative examples. In recent

[*] We would like to thank Sam Hellmuth and two anonymous reviewers for their valuable comments. This paper is part of the projects D1 "Linguistic Databases for Information Structure: Annotation and Retrieval" and D2 "Typology of Information Structure" at the University of Potsdam (sponsored by the German Research Foundation).

B.D. ten Cate and H.W. Zeevat (Eds.): TbiLLC 2005, LNAI 4363, pp. 127 – 138, 2007.

years, there have been some attempts to create typological archives containing primary data (texts and sound files). A well-known example is the *LACITO Archive*[1], which contains texts or single sentences collected from Oceanic languages, Caucasian languages, and languages of Nepal. This is a new and promising type of resource for typological studies, that combines the properties of a typological database and the properties of a natural language corpus containing a large collection of primary language data.

In this paper we present a contribution to the development of typological archives of this type. Two properties of the resource we are presenting in this paper are innovative with respect to previous attempts: first, the data from the different languages is collected by identical data collection methods, i.e., the data set is a type of a parallel corpus, and not just a resource for the archiving of data from more than one language; second, the data is richly annotated with a large number of linguistic layers (phonology, morphology, syntax, semantics, and information structure), hence allowing the user to explore the occurrence of grammatical categories in the entire set of archived data.

The aim of this paper is to present the structure of this resource along with the means of exploring it. In Section 2, the data contained in the database is described in more detail. In Section 3, we discuss the requirements of an exploration interface and present our current solutions, and Section 4 summarizes the main points of this article.

2 A Cross-Linguistic Production Data Archive

The cross-linguistic empirical data is collected using the *Questionnaire on Information Structure* (henceforth, QUIS, see [13]).[2] The aim of this tool is to provide methods for the collection of data for the study of information structure (henceforth, IS) in the object language. QUIS comprises a set of translation tasks and production experiments for the collection of primary data (see Sect. 2.2). The "translation tasks" contain a number of simple sentences in particular contexts and question/answer pairs illustrating a range of IS categories that are translated into the object language from a contact language. The "production experiments" contain a range of experimental settings that induce spontaneous expressions (e.g., picture descriptions, map tasks, etc.; see details in Sect. 2.2). Finally, QUIS provides a section with questions about the grammatical structure of the language (see Sect. 2.3).

On the basis of these data collection methods, a corpus of primary data is currently being built up from fifteen languages belonging to different language families and spoken in different parts of the world, with about 2,000 sentences per language:

[1] See http://lacito.vjf.fr/archivage/index.html.

[2] The cross-linguistic production data archive is being developed within the Sonderforschungsbereich 632 "Information Structure" at the University of Potsdam and the Humboldt University Berlin funded by the German Scientific Society (DFG). The development of this archive is collaborative work of the projects D1 "Linguistic Database for Information Structure: Annotation and Retrieval" (S. Dipper, M. Götze, J. Ritz, M. Stede), and the project D2 "Typology of Information Structure" (G. Fanselow, C. Féry, I. Fiedler, M. Krifka, A. Schwarz, R. Stoel, and S. Skopeteas). For more information, see http://www.sfb632.uni-potsdam.de.

Chinese, French, Dutch, Georgian, German, Greek, English, Hungarian, Japanese, Konkani (India: Indo-Iranian), Maung (Australia: Non-Pama-Nyungan), Niue (Niue: Austronesian), Prinmi (China: Tibeto-Burman), Teribe (Panama: Chibchan), and Yucatec Maya (Mexico: Mayan). For every collected sentence, the database contains the following:

(a) sound file;
(b) transcription;
(c) annotation;
(d) metadata.

Besides the data from the individual languages, the database contains full documentation of the experiments and translation tasks and further supporting documents concerning the performance of the experiments and the archiving methods.

2.1 Primary Data and Annotation

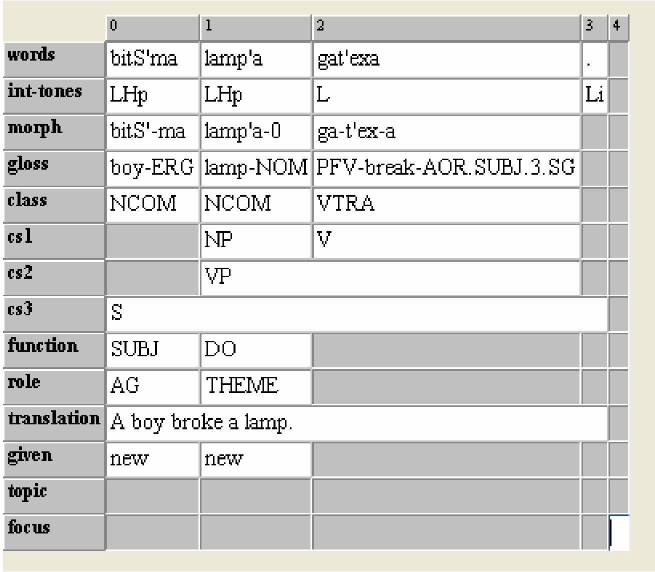
The primary data is collected in the place where each language is spoken. Translation tasks and production experiments are performed by native speakers under the guidance of researchers specialized in the grammatical description of the object language. The data is recorded in the field then digitized and prepared for insertion in the database using *Praat* (see [3]).

The sound files are transcribed and annotated using *EXMARaLDA* (see [20]). The annotation is based on detailed annotation guidelines (see [10]), with an annotation scheme providing a comprehensive description on the following layers: phonological (orthographic and phonemic transcription, lexical tones, intonational tones, breaks, and prosodic structure, as well as further optional features), morphological (morphemic transcription, glossing, and word class), syntactic (grammatical functions, semantic roles, and constituent structure), semantic (free translation, definiteness, countability, animacy, and quantificational properties) and information structural annotations (givenness, topic, and focus). The development of the detailed annotation guidelines is the collaborative product of interdisciplinary working groups in which researchers of different projects of the Collaborative Research Center participated.

The annotation files are illustrated in Fig. 1 by means of a Georgian sentence (screenshot from the *EXMARaLDA* editor). In Fig. 1, only a part of the annotation tiers is displayed for illustrative purposes. The tier *words* contains a phonological transcription of the spoken utterance and the tier *int-tones* indicates in auto-segmental-metrical notation the tonal events that accompany it. The utterance is morphologically transcribed in *morph*, which indicates morpheme boundaries. The tier *gloss* presents a morpheme-to-morpheme translation following the glossing conventions established in typological studies (see *Eurotyp* in [17] or *LGR* in [1]) and the tier *class* contains information about the word class (the abbreviations follow the general conventions established in EAGLES[3]). Subsequent tiers describe the syntactic properties of the utterance: cs_n represent the constituent structure, *function* and *role* provide information about syntactic function and semantic role, respectively. After

[3] http://www.ilc.cnr.it/EAGLES96/browse.html

the free translation of the example, the last three tiers illustrate the annotation of information structure: since the example has been elicited in an out-of-the blue context, all referential nominal phrases (NPs) bear new information.



	0	1	2	3	4
words	bitS'ma	lamp'a	gat'exa	.	
int-tones	LHp	LHp	L	Li	
morph	bitS'-ma	lamp'a-0	ga-t'ex-a		
gloss	boy-ERG	lamp-NOM	PFV-break-AOR.SUBJ.3.SG		
class	NCOM	NCOM	VTRA		
cs1		NP	V		
cs2		VP			
cs3	S				
function	SUBJ	DO			
role	AG	THEME			
translation	A boy broke a lamp.				
given	new	new			
topic					
focus					

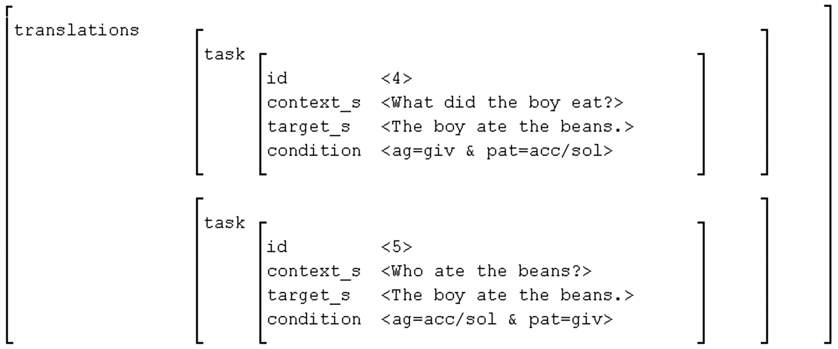
Fig. 1. Annotated expression (Georgian, annotated by R. Asatiani)

2.2 Data Gathering Methods

This Section presents the data gathering methods that are used in QUIS: translation tasks and production experiments.

Elicitation through translation is a commonly used method for data collection, especially in cross-linguistic comparison (see [7]). Following this research tradition, QUIS contains 252 simple discourse units that are given in English and are translated and recorded by native speakers in the object languages (when necessary through the medium of a further contact language). These discourse units contain a target sentence often preceded by a context sentence (either a question or a declarative). The context is used to manipulate the discourse condition in which the target sentence is produced, hence evoking information structural effects on it. For instance, the sentence *The boy ate the beans* is translated and recorded as an answer to the questions: (a) *What did the boy eat?* and (b) *Who ate the beans?* Depending on the object language, the context questions may trigger different syntactic, morphological and/or prosodic structures in the answer. Further translation tasks are used to induce several types of topic and focus or manipulations of the discourse status of the referents.

The translation tasks are labeled for the discourse conditions in which the target sentence is assumed to be realized. So, the context question presented in translation task “4” in Fig. 2 evokes the discourse condition “the agent is given and the theme is solicited through the question”. The definite expression of the theme in the target sentence requires that the theme is accessible information for the discourse participants. Translation task “5” is designed to evoke the reverse discourse conditions for the same target sentence.



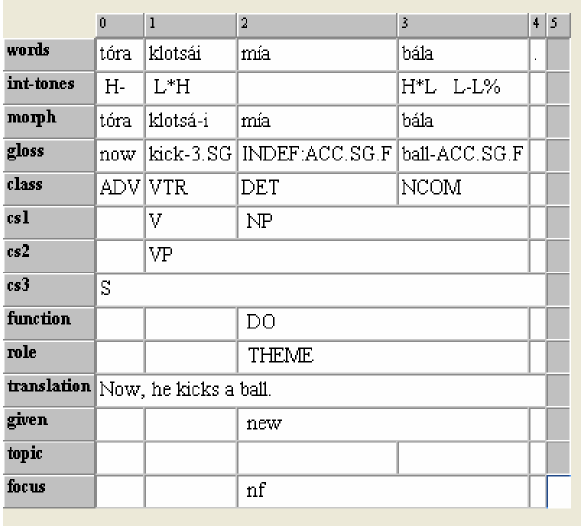
```
translations  task  id         <4>
                    context_s  <What did the boy eat?>
                    target_s   <The boy ate the beans.>
                    condition  <ag=giv & pat=acc/sol>

              task  id         <5>
                    context_s  <Who ate the beans?>
                    target_s   <The boy ate the beans.>
                    condition  <ag=acc/sol & pat=giv>
```

Fig. 2. Structure of the translation tasks

QUIS also contains 30 production experiments which all make use of visual stimuli (either pictures or films), so that the data from the different languages is induced by a cross-linguistically invariant perceptual input. Each experiment aims to compare among different discourse conditions which are established through the stimuli and the experimental instruction. Each condition is factorially implemented with a set of different stimuli (for example, different pictures that correspond to different events), in order to ensure that the resulting observations are not influenced by event- or item-particular effects. Depending on the experimental design, the production experiments are performed by four to eight native speakers, who each see the same items but in different conditions.

As an illustrative example we will discuss a production experiment that is intended to induce manipulations of the discourse status of the arguments through the description of picture sequences. The picture sequences implement several discourse conditions of which two are described here. Condition A is intended to induce the production of a sentence in which the agent is given information and the theme is new. In order to achieve this discourse condition, the first picture of the sequence (context situation) presents an entity *x*, e.g. “a man”, and the second picture (target situation) presents an event, in which entity *x* is involved as an agent and a new entity *y* is involved as a theme, e.g., “the man is kicking a ball”. The data from Condition A is compared with the data from Condition B, which is intended to induce the production of a sentence in which the agent is new information and the theme is given. In order to induce this information structure, the first picture of a sequence (context situation) presents the entity that is involved as a theme in the event of the second picture (target situation), e.g. picture 1 presents “a ball” and picture 2 presents “a man kicking the ball”. The native speakers are shown the pictures one after the other and are instructed to describe the situations which are presented to them as a coherent story. The data gathered through this experiment might allow for generalizations concerning the use of pronouns, the use of different word orders, and the occurrence of active/passive voice in the object languages.

The following examples illustrate the kind of data that are obtained through production experiments and their annotations. Condition A of the experiment under discussion induced in Modern Greek the target sentence shown in Fig. 3. The given agent is not encoded through a lexical NP, but is cross-referenced by the subject suffix on the verb. The new theme is encoded through an indefinite NP. Only the overtly encoded referents are annotated in the layer of information structure: The object constituent is annotated as *new* (see the givenness tier, labelled “*given*” in the leftmost column).

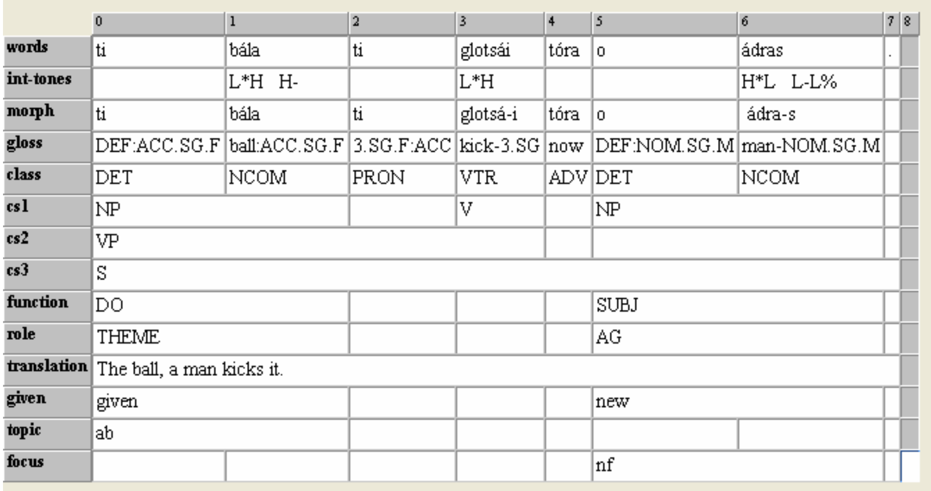


	0	1	2	3	4	5
words	tóra	klotsái	mía	bála	.	
int-tones	H-	L*H		H*L L-L%		
morph	tóra	klotsá-i	mía	bála		
gloss	now	kick-3.SG	INDEF:ACC.SG.F	ball-ACC.SG.F		
class	ADV	VTR	DET	NCOM		
cs1		V	NP			
cs2		VP				
cs3	S					
function			DO			
role			THEME			
translation	Now, he kicks a ball.					
given			new			
topic						
focus			nf			

Fig. 3. Target sentence in Condition A (Modern Greek)

Condition B is illustrated in Fig. 4. The given theme is left dislocated in this example; it is annotated as *given* in the givenness tier (label “*given*”) and as an aboutness topic (*ab*) in the tier *topic*. The new agent is encoded through the postverbal subject NP.

Data gathered through production experiments contains the spontaneous reactions of native speakers. In consequence, the structure that the native speaker produces during the performance of the experiment often deviates from the predicted structure. The example in Fig. 4 illustrates a deviation of this kind. Although the agent ‘the man’ is a new referent (i.e., not mentioned in the previous discourse), it is encoded as a definite NP in the illustrated example. This is captured through the annotation: the gloss shows that the native speaker has used a definite article, but the givenness tier (label "*given*") shows that this constituent is new information.



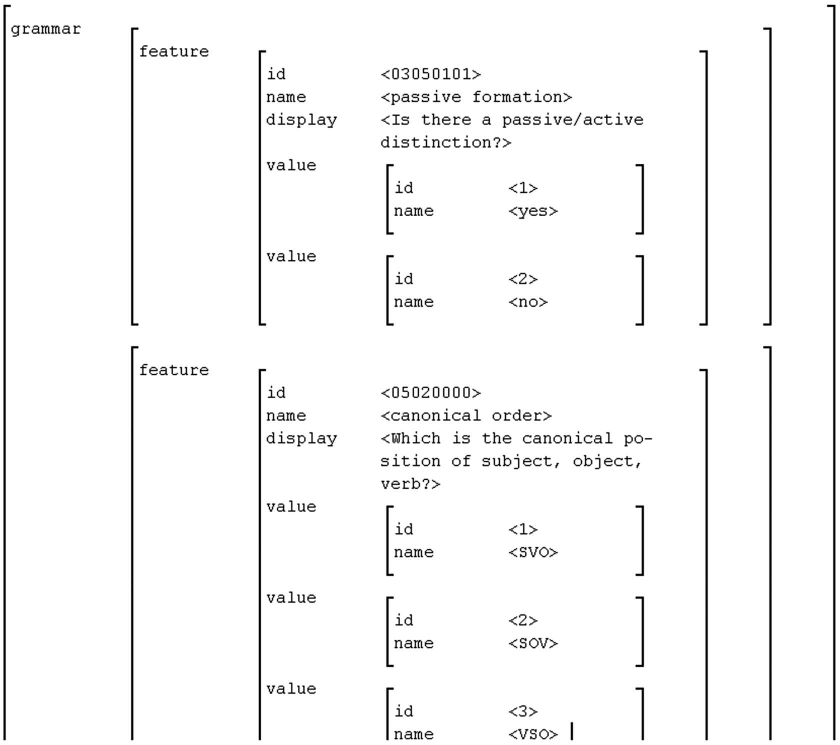
	0	1	2	3	4	5	6	7	8
words	ti	bála	ti	glotsái	tóra	o	ádras	.	
int-tones		L*H H-		L*H			H*L L-L%		
morph	ti	bála	ti	glotsá-i	tóra	o	ádra-s		
gloss	DEF:ACC.SG.F	ball:ACC.SG.F	3.SG.F:ACC	kick-3.SG	now	DEF:NOM.SG.M	man-NOM.SG.M		
class	DET	NCOM	PRON	VTR	ADV	DET	NCOM		
cs1	NP			V		NP			
cs2	VP								
cs3	S								
function	DO					SUBJ			
role	THEME					AG			
translation	The ball, a man kicks it.								
given	given					new			
topic	ab								
focus						nf			

Fig. 4. Target sentence in Condition B (Modern Greek)

Besides simple picture descriptions, the production experiments of QUIS include several types of tasks, such as map tasks, spontaneous answers to questions, instruction games between two informants (e.g., an informant gives instructions to the other for the development of a spatial configuration), role games (e.g., two informants see a short film and perform a negotiation), etc.

2.3 General Questions on the Grammar

This component of the cross-linguistic production data archive relies on the tradition of typological questionnaires (see [5]) and has the structure of a typological feature database such as those mentioned in Section 1. It contains several questions on the typological properties of the grammar (phonology, morphology, syntax, and information structure) of the object language, that are necessary for the interpretation of the collected data. Each section contains a number of grammatical features that are presented to the user as questions, e.g. "Is there a passive/active distinction?", or "what is the canonical position of subject, object, verb?". The fragment in Fig. 5 presents the hierarchical structure of this component in the database. Each feature is accompanied by a finite set of values, that represent the typologically possible options: For the first example ("Is there a passive/active distinction?"), the possible options are "yes" and "no"; for the second example the possible options are the word orders encountered in world's languages: "SOV", "SVO", "VSO", "VOS", "OSV", and "OVS". The answers for these questions are not inferred from the archived data, but are collected from available grammatical descriptions and from the grammatical knowledge of language experts.



```
grammar  feature  id        <03050101>
                  name      <passive formation>
                  display   <Is there a passive/active
                            distinction?>
                  value     id      <1>
                            name    <yes>
                  value     id      <2>
                            name    <no>

         feature  id        <05020000>
                  name      <canonical order>
                  display   <Which is the canonical po-
                            sition of subject, object,
                            verb?>
                  value     id      <1>
                            name    <SVO>
                  value     id      <2>
                            name    <SOV>
                  value     id      <3>
                            name    <VSO> |
```

Fig. 5. Fragment from the grammatical data

The grammatical information contained in this component is indispensable for the interpretation of the production data. For instance, the data gathered in the condition "agent=new information" in the experiment presented above contains a large number of passive sentences in languages like English and German, but only active sentences in the data from Prinmi and Georgian.[4] Crucially for the interpretation of the result, the grammar of Prinmi does not have a passive formation rule (see [8]), while passive formation is available for the Georgian verb (see [14]). I.e., in these two languages the same experimental result is observed, but for completely different reasons: in Prinmi, passive is not an available option, whereas in Georgian passive is available but is not chosen in the discourse condition at issue. This information about the grammar is given through the values of Georgian ("yes") and Prinmi ("no") in the feature "passive formation" of Fig. 5.

3 Exploring the Archive

In sum, our production data archive differs from other typological data archives in providing: (a) primary data with rich multilevel annotations, (b) information about the

[4] The production data archive contains 16 sentences in each language gathered through this experimental condition (produced by eight different speakers per language).

discourse condition that induces the archived data, and (c) grammatical information about the object language. In this section, we present the solutions we have chosen in order to develop the production data archive and we illustrate the possibilities currently available for exploration of the archive. We do not illustrate in detail how searches are performed within the annotated data, since these do not substantially differ from exploration in text corpora (the reader is referred to [12] for natural language data and annotation instead), but we focus on the possibilities that emerge from the integration of the various components presented above into a single exploration environment.

As representation formats for individual archive components, we employ both existing formats and formats developed within the framework of the *Collaborative Research Center 632 "Information Structure"*. For representation of the natural language data and their annotations, the generic standoff XML exchange format PAULA is used, which facilitates easy addition of further annotation layers and supports import from a number of annotation tool formats (ref. [9]). For accessing of this data, we use ANNIS, a web application that allows for visualization and querying of the heterogeneous multilevel annotation via the internet ([11]).

We are currently developing an XML-format for QUIS, in particular for the grammatical questionnaire and the documentation of the data gathering methods. For visualization and querying of the Questionnaire, we are developing an exploration system, which will integrate as an interface to ANNIS, such that the production data archive can be viewed in a single environment.

An elementary way of searching within the production data archive is to query the annotations. For this purpose, the user of the archive may formulate query expressions that address any aspect of the information that is archived within the production data archive. A standard query would retrieve all sentences of a given language that match certain properties in the annotation, e.g.: "For the language with the name *Teribe* (TFR), retrieve sentences in which the agent (tag *ag* of type *role*) is the part of the sentence that constitutes the answer to a previous question (tag *ans* of type *focus*)".

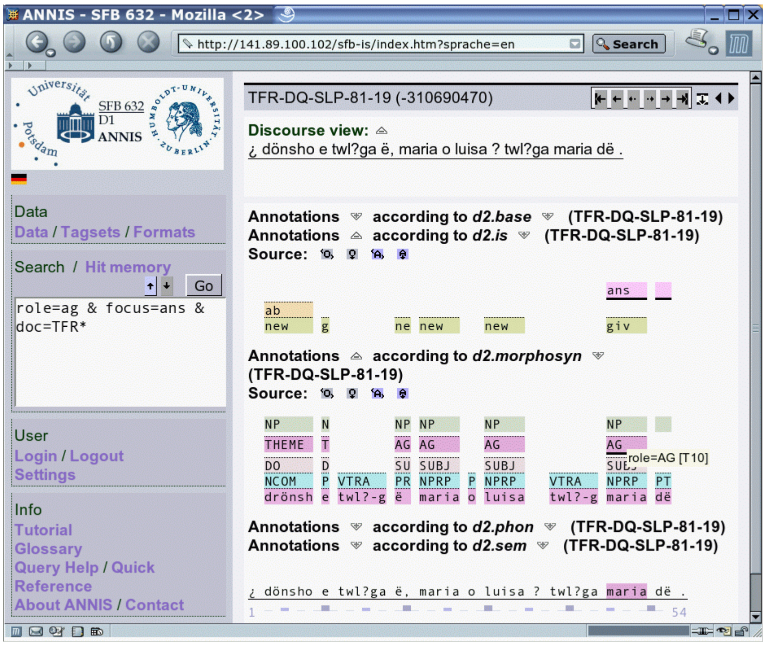
role=ag & focus=ans & doc=TFR* (1)[5]

The result of the query in (3) is shown in the screen-shot from ANNIS in Fig. 6:

As discussed in Section 2, a powerful property of the production data archive is that it not only provides an annotated text corpus, but also a description of the discourse environments (experimental conditions) in which this expression was induced. Experiments and experimental conditions are specified through the file names. In Section 2.2, for example, we have shown that an agent that is assumed to be new according to the experimental manipulation may be encoded through definite NPs in the resulting data. In order to retrieve examples like the one in Fig. 4, the user may address a particular experimental condition, e.g.: "For the language with the name *Greek* (GRK), retrieve sentences gathered in experiment 42, condition B, in which the agent (tag *ag* of type *role*) is encoded through an NP that contains a definite article (tag *def* of type *gloss*)."

role=ag & gloss=def & doc=GRK*42-B & #1_=_#2 (2)

[5] The expressions "doc=TFR" and "doc=GRK" (below) restrict the search to documents of Teribe (language code: TFR) or Greek (language code: GRK).



```
role=ag & focus=ans &
doc=TFR*
```

TFR-DQ-SLP-81-19 (-310690470)

Discourse view:
¿ dönsho e twl?ga ë, maria o luisa ? twl?ga maria dë .

Annotations according to *d2.base* (TFR-DQ-SLP-81-19)
Annotations according to *d2.is* (TFR-DQ-SLP-81-19)
Source:

```
                                                ans
ab
new       g       ne  new     new               giv
```

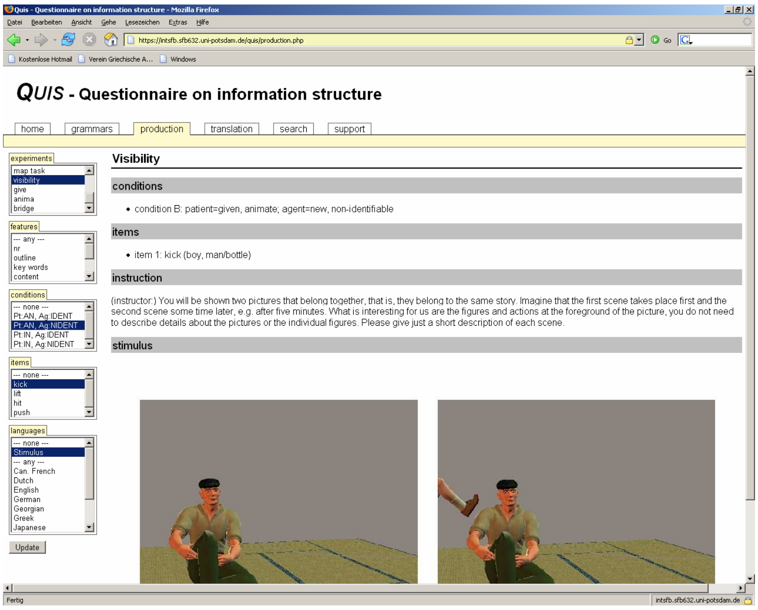
Annotations according to *d2.morphosyn* (TFR-DQ-SLP-81-19)
Source:

```
NP      N       NP  NP      NP            NP    
THEME   T       AG  AG      AG            AG
                                              role=AG [T10]
DO      D       SU  SUBJ    SUBJ          SUBJ
NCOM    P VTRA  PR  NPRP  P NPRP   VTRA   NPRP  PT
drönsh  e twl?-g ë  maria o luisa  twl?-g maria dë
```

Annotations according to *d2.phon* (TFR-DQ-SLP-81-19)
Annotations according to *d2.sem* (TFR-DQ-SLP-81-19)

¿ dönsho e twl?ga ë, maria o luisa ? twl?ga maria dë .

Fig. 6. ANNIS



QUIS - Questionnaire on information structure

Visibility

conditions

- condition B: patient=given, animate; agent=new, non-identifiable

items

- item 1: kick (boy, man/bottle)

instruction

(instructor:) You will be shown two pictures that belong together, that is, they belong to the same story. Imagine that the first scene takes place first and the second scene some time later, e.g. after five minutes. What is interesting for us are the figures and actions at the foreground of the picture, you do not need to describe details about the pictures or the individual figures. Please give just a short description of each scene.

stimulus

Fig. 7. QUISViewer

For browsing of the documentation of the experiments in QUIS, we are currently developing a tool "QUISViewer". The documentation includes an outline of each experiment and its experimental conditions, the procedure and instructions that were used during performance of the experiment, as well as the stimuli which were shown to the informant. A pilot version of this browser is shown in Fig. 7. At the left frame of this interface, the user may also browse the collected data restricting his query to particular experimental conditions, experimental items, or a subset of languages.

The aim of the component of QUIS which provides general questions on the grammar (see Sect. 2.3) is to support typological queries within our archive. Currently, the information about grammatical features of each languages is available in a separate database. In a future development this information will be integrated into ANNIS to allow for queries of the type: "For a language L_i such that it has either value 'VOS' or 'VSO' in the feature 'canonical order', retrieve sentences in which a noun phrase precedes a verb".

4 Summary

We have presented our work on a cross-linguistic production data archive, which includes detailed information about data collection methods and about grammatical features of the languages involved, in addition to richly annotated natural language data from 15 typologically diverse languages. The special property of our archive is that it contains a parallel corpus of sentences and texts induced in the different languages through identical methods. We argue that this is a new type of resource that integrates features from both typological databases and natural language corpora. Finally, we have sketched the possibilities available for exploration of this archive on the basis of our current implementation, emphasizing operations that take place at the interfaces between the database components, in order to give an insight into the special properties of our complex archive architecture.

We believe that the type of resource presented in this paper represents a substantial enrichment of existing resources for language comparison, since it permits formulation of generalizations about the occurrence of language specific patterns in identical conditions. The *Collaborative Research Center 632 "Information Structure"* plans in future to expand the database with data from additional languages. Parallel to the integration of further data, we will also further develop the archiving infrastructure towards an integrated environment containing all of the components reported in this paper.

References

1. Bickel, B., Comrie, B., Haspelmath, M.: Leipzig Glossing Rules. Ms. University of Leipzig (2004)
2. Bickel, B., Nichols, J.: Autotypologizing Databases and their Use in Field Work. In: Proc. Int. LREC Workshop on Resources and Tools in Field Linguistics (2002)
3. Boersma, P., Weenink, D.: Praat. doing phonetics by computer (Version 4.3.14) (2005), Computer program: http://www.praat.org/

4. Brown, D., Corbett, C., Tiberius, C., Barron, J.: The Surrey Database of Agreement (2005), Online database: http://www.smg.surrey.ac.uk/Agreement/explore.aspx
5. Comrie, B., Smith, N.: Lingua Descriptive Studies: Questionnaire. Lingua 42, 1–72 (1977)
6. Corbett, C., Baerman, M., Brown, D., Hippisley, A.: Extended Deponency: The Right Morphology in the Wrong Place (2005), Online database: http://www.surrey.ac.uk/ LIS/ MB/WALS/WALS.htm
7. Dahl, Ö. (ed.): Tense and Aspect in the Languages of Europe. Mouton de Gruyter, Berlin, New York (2000)
8. Ding, S.: Fundamentals of Prinmi. A Tibeto-Burman Language of Northwestern Yunnan, China. PhD. dissertation, Australian National University (1998)
9. Dipper, S.: XML-Based Stand-off Representation and Exploitation of Multi-Level Linguistic Annotation. In: BXML 2005. Proceedings of Berliner XML Tage 2005, Berlin, pp. 39–50 (2005)
10. Annotation Guidelines. In: Dipper, S., Götze, M., Skopeteas, S. (eds.) Interdisciplinary Studies on Information Structure (ISIS). Working Papers of the SFB 632, vol. 8, Universitätsverlag Potsdam, Potsdam (2006)
11. Dipper, S., Götze, M., Stede, M., Wegst, T.: ANNIS. A Linguistic Database For Exploring Information Structure. In: Interdisciplinary Studies on Information Structure (ISIS). Working Papers of the SFB 632, pp. 245–279. Universitätsverlag Potsdam, Potsdam (2004)
12. Dybkjaer, L., Berman, S., Bernsen, N.O., Carletta, J., Heid, U., LListerri, J.: Requirements Specification for a Tool in Support of Annotation of Natural Interaction and Multimodal Datad. ISLE Natural Interactivity and Multimodality Working Group. D11.2 (2001)
13. Skopeteas, S., Fiedler, M., Hellmuth, I., Schwarz, S., Stoel, A., Fanselow, R., Féry, G., Krifka, C.: Questionnaire on Information Structure. In: Interdisciplinary Studies on Information Structure (ISIS). Working Papers of the SFB 632, vol. 6, Universitätsverlag Potsdam, Potsdam (2006)
14. Harris, A.C.: Georgian Syntax. Cambridge University Press, Cambridge (1981)
15. Haspelmath, M., Dryer, M.S., Gil, D., Comrie, B. (eds.): The World Atlas of Language Structures. Oxford University Press, Oxford (2005)
16. Hyman, L., Mortensen, D., Allison, D.: X-tone: Cross-linguistic Tonal Database (2005), Online database: http://xtone.linguistics.berkeley.edu/display/index.php
17. König, E., Bakker, D., Dahl, Ö., Haspelmath, M., Koptjevskaja-Tamm, M., Lehmann, C., Siewierska, A.: EUROTYP Guidelines. European Science Foundation Programme in Language Typology (1993)
18. König, E., Gast, V., Hole, D., Siemund, P., Töpper, S.: Typological Database of Intensifiers and Reflexives. Freie Universität Berlin (2006), Online Database: http:// noam. philologie.fu-berlin.de/ gast/tdir/
19. Hurch, B., Mattes, V.: The Graz Database on Reduplication. Faits de Langues (to appear)
20. Schmidt, T.: Transcribing and Annotating Spoken Language with EXMARaLDA. In: Proceedings of the LREC-Workshop on XML Based Richly Annotated Corpora, Lisbon 2004. ELRA, Paris (2004)
21. Wittenburg, P., Mosel, U., Dwyer, A.: Methods of Language Documentation in the DOBES Project. In: Proceedings of LREC 2002, pp. 34–42 (2002)