University of Dublin
Trinity College Dublin
School of Linguistic, Speech and Communication Sciences
Centre for Language and Communication Studies
LI 7864 Corpus Linguistics
Dr Elaine Uí Dhonnchadha
Michaelmas Term 2009

10.01.2010

The Design and Implementation of Multimodal Corpora

The AIBO Corpus

by
Daniel Jettka
jettkad@tcd.ie

Content

FIGURES

1 Introduction

This essay deals with the basic principles of the design and implementation of multimodal corpora. The introductory theoretical assumptions made in the course of the essay will be complemented by the presentation of pratical experiences which originate from the creation of the AIBO corpus. This multimodal corpus, focussing on the interaction between German students and the SONY robot dog AIBO, was established in 2006 in the context of the seminar „Methods for Representing and Processing Multimodal Documents“ which was held at the University of Bielefeld under the direction of Prof. Dr. Dieter Metzing.

Starting with a theoretical introduction, chapter 2 will give an overview of the building of multimodal corpora. In specific, the recording of video data for the AIBO corpus and its administration will be demonstrated from a theoretical as well as practical point of view.

After the initial steps, the participating student groups, consisting of three to four students, chose a few characteristic video sequences in order to build a small annotated corpus. The necessary next step was the operationalisation of the data by the execution of annotations. In this context, the annotation tool ELAN seemed useful and promising, since it allows for the easy and structured annotation of video and audio data. Therefore, it represents an ideal starting point for the annotation and analysis of multimodal interaction. In order to annotate data by ELAN, annotation schemes for the underlying modalities had to be defined. Amongst the modalities, which were considered to be of interest, were speech, facial expression, gesture, and the LED signals of AIBO. The relations of the modalities, annotation schemes and ELAN will be discussed in chapter 3.

The interrelations of the individual modalities could be analysed on the basis of the video data. There are different strategies for the analysis of corpora. These can be divided into qualitative and quantitative approaches, whereas both of them include

certain advantages and disadvantages. Certainly, the quantitative approach is the standard method for analysing corpora, but because of a lack of time, it could not be realised for the AIBO corpus. Nevertheless, chapter 4 will present the basic notions of both, quantitative and qualitative, approaches towards corpora.

Finally, the concluding remarks (chapter 5) will recapitulate the basic characteristics of multimodal corpora. Furthermore, there will be a discussion of the problems which generally occur in the context of multimodal corpora and those which specifically were discovered during the design and implementation of the AIBO corpus.

2 Building a multimodal corpus

2.1 What is a multimodal corpus and why do we need it?

In order to define the basic characteristics of a multimodal corpus, at first, one could have a look at the two individual parts of the term: "multimodal" and "corpus". A corpus in linguistics is meant to refer to a set of language data of similar type, e.g. texts. In modern linguistics such sets normally exist in the form of digitalised, computer-readable data. This allows for effective searchablity and further processing (cf. McEnery, 2008, p. 31f.).

There are other fields of interest than texts, for instance human interaction, which includes various modalities like speech or lip, eye and head movements, body postures, gestures, hand shapes, facial expressions, and haptics (cf. Wittenburg, 2008, p. 664). The consideration of several or at least two of such modalities is a condition for making a corpus a multimodal corpus, i.e. the language data in a multimodal corpus has to allow for the analysis of more than one modality. This does not mean that corpora consisting of texts cannot focus on several descriptive levels, since one could imagine a corpus of handwritten texts, e.g. considering the textual content and the shape of the handwriting, or texts being analysed on several linguistic levels (orthography, syntax, discourse structure, etc.). However, this should be differentiated from the general notion of multimodal corpora, which consider several medially separated modalities. Hence, it would be useful to distinguish multimodal and multidimensional corpora.

Certainly the most widespread application of multimodal corpora is the examination of human interaction, which is also a part of the AIBO corpus that captures human-robot as well as human-human interaction. Accordingly, the main purpose of multimodal corpora lies in the recording and analysis of human interaction and the "construction of meaning and understanding" (Allwood, 2008, p. 210f.). This would

not be possible by language data which does not capture the involved modalities. Furthermore, in opposition to the "acceptability judgments or grammaticality judgments" (Gries, 2006, p. 191) in earlier linguistic research, the analysis of interaction based on corpora, because of their empirical and quantitative orientation, generally allows for more exact results with respect to objectivity, reliability, versatility and validity (cf. ibid.).

There is no specific size that makes a set of language data a corpus, but a general assumption is that corpora have a finite size (cf. McEnery, 2008, p. 30f.). This can be different for so-called "monitor corpora" (ibid.) which have a special diachronic focus on linguistic subfields, e.g. the lexicographic view in case of the COBUILD group (http://www.collins.co.uk/Corpus/CorpusSearch.aspx).

2.2 Recording and administration

Since this essay focusses upon corpora regarding human interaction, the more specific method of recording (as opposed to the more general 'collecting' in other contexts) will be presented here. There are mainly two questions for the recording of data for building a multimodal corpus: what to record and how to record.

The first issue mostly depends on the goal of the corpus study, namely what is to be examined and where it is assumed to be found. There are various factors which have to be addressed and may be controlled before starting to record data: the participants of the study and their social-structural features, the context of the recording, i.e. the environment, its influence on the data as well as the topic of the recorded interaction. All of these aspects can have an impact on the recorded data and hence on the results of a subsequent analysis.

The second issue, how to record the data, addresses the method of recording and the use of particular recording devices. The choice of equipment likewise depends on

the domain of the study and is a factor for the analysis of the data. Whereas in classical anthropological field studies written notes were used (cf. Allwood, 2008, p. 214), in modern times more technically specialised methods come into play. Depending on the modalities taken into consideration, special devices like ultrasonic or infrared emitters (gestures), data gloves (hand shape), eye-trackers (eye movement), light reflectors (body movement), etc. are available (cf. Wittenburg, 2008, p. 667). Their use naturally is up to the budget and the specific focus of the project. Additionally, it must be emphasised that their use can have impacts on the naturality of the interaction, e.g. by alternating certain behaviour simply by their presence.

This is one reason, besides the very striking financial aspect, for the use of video cameras as more holistic recording devices. A problem resulting in the more holistic attempt is to capture as much information as possible, at least the one which will be important for the analysis of the interaction. Therefore, the use of several devices can be advisable, e.g. microphones for exact audio information, cameras focussing special aspects, like the current speaker, cameras with an overall view or two normal 2D cameras which can be used to reconstruct the 3D space. This in turn rises the requirement of synchronisation of the individual devices which can be realised technically either during the recording or semi-automatically afterwards (cf. Wittenburg, 2008, p. 665).

The video recordings for the AIBO corpus were meant to capture the interaction between the involved humans as well as the interaction between humans and the Sony robot dog AIBO (http://www.sonydigital-link.com/AIBO/). The basic arrangement for the recordings was given by the following situation:

Figure 1: Initial situation of video recording

Two persons from a student group had to freely interact with AIBO, without any particular instructions. Their only equipment was an overview of the most important speech commands for AIBO (cf. Appendix E, p. 28). The particular situation shown in Figure 1 also included the owner of AIBO who gave a short introduction into its functions.

The recordings were made by two fixed video cameras and one handheld camera which was carried by a non-participating person of the group. The camera man/ woman was the only person in the room who was assumed not to be directly involved in the recorded interaction. By using three cameras, the interaction should be recorded as detailed as possible. This strategy is eminently important to capture all aspects of the action, even those which prove to be interesting retrospectively. One group became especially conscious of this problem after the recordings, realising that the video cameras did not record the LED signals of AIBO's display (cf. section 3.1.4, p. 13), which turned out to be very important for their analysis. This emphasises the necessity for the capturing of as many details as possible.

The recordings (12 x ca. 30 minutes) were guided by the team of the Bielefeld University multimedia laboratory and were sent to the students in .mov and .mpeg format plus additional .wav audio track. The choice of format was a conscious decision since “MPEG2 is seen as the backend video format for archiving and digital libraries, and it is supported by a wide range of software products” (Wittenburg, 2008, p. 669).

3 Annotation

Annotating a corpus can be seen as giving "added value" (Leech, 2005) to it. There is no consensus in research if an annotation can be seen as an integral part of a corpus because every annotation necessarily reflects an interpretative action of the annotator, e.g. by choosing a specific annotation scheme or making certain decisions in doubtful cases. Hence, it is not completely possible to logically separate the analysis from the annotation of a corpus. Therefore, the crucial distinction of raw corpora and annotated corpora should be made.

Despite of this, an annotation can be seen as a very important step in the building, processing or analysis of a corpus; accordingly, where to logically settle it, depends on the theoretical point of view. It engages the automatic query of corpora, the relating of certain aspects from distinct modalities, more advanced automatic analyses (e.g. syntactic analysis, anaphora resolution, etc.) as well as the re-usability for later studies or those with a different research focus (cf. Leech, 2005).

There sholuld be some important considerations before starting to annotate the raw corpus, e.g. if the annotation should be carried out manually, semi-automatically or even completely automatically. Another point is if the annotation should cover all or just parts of the data. These considerations mostly depend on the field of interest of the researcher, i.e. which hypotheses were being made and how are they thought to be tested (and of course money and time issues again).

Besides, there are some definite requirements for annotations (cf. Leech, 2005). Firstly, a comprehensive documentation should be given, including a wide range of metadata, i.e. data about the data (participants, time data, situation, annotator, etc.), and an accurate description of annotation schemes (cf. section 3.1). Secondly, a synchronisation of recording and annotation has to be assured. That means that both of these have to refer, as exactly as possible, to the same point in time or time frame, which is normally no problem when using annotation tools for multimodal data (cf.

section 3.2). Thirdly, the separation of primary data (raw recording) and annotation has to be guaranteed: there must not be a direct manipulation of data, for instance by drawing arrows directly into the single images of the video data. This normally is respected by annotation tools, but it has to be emphasised since the ability to reconstruct the primary data can be of great importance, e.g. for using it in other contexts or correcting errors at any time.

According to the previous assumptions, the next step of the AIBO project was to view the recorded data and extract interesting aspects. In order to make an annotation realisable in the context of the seminar, 14 short sequences (ranging from 10sec up to 2:20min) were chosen, which seemed especially interesting for analysing the human-robot interaction. The main tasks for the ongoing analysis of the data mainly consisted in (a) fixing the observations by assigning explicitly defined categories and (b) testing the hypotheses with regard to the categories.

3.1 Annotation schemes

There are some basic theoretical requirements for annotation schemes which should be taken into account. First of all, there should be a detailed documentation including a precise idea of the phenomenon to be captured, which is a necessary precondition for the definition of adequate categories.

In this context, it is advisable to initially orient towards present “de facto standards”. These can be seen as “some kind of standardisation that has already begun to take place, due to influential precedents or practical initiatives in the research community” (Leech, 2005). If applicable, these have certain advantages in comparison to self-defined categories. Not only do they save time and money, but primarily they make

an annotation much more sustainable and re-usable (cf. ibid.). The following sections will show that the creation of the AIBO corpus respected this premise.

A basic characteristic of annotation schemes is the definition of a fixed set of categories which are more or less interpretatively used for the description of language data. This could be seen as a disadvantage since it eminently constrains the specificity of the annotation. Concurrently, this is a necessary step in allowing for the reduction of erroneous annotations, individually and between separate annotators (increase of inter-annotator agreement). Moreover, it can be seen as a precondition for the statistical analysis of annotations, for which, in the case of free annotations, there could be problems to identify equally or similarly annotated segments, inasmuch as annotations could differ in their exact wording.

The following sections will cover the annotation schemes which formed the basis of the annotation of the AIBO corpus. These schemes fulfill the duty to define a frame of fixed categories, in order to capture the multimodal aspects of the recorded interaction and to allow for a high degree of reliability of the annotation. Furthermore, the data can be used in long term and later studies since the clearly defined annotation schemes can be transformed into other formats much more easily than free annotations.

3.1.1 Speech: SAMPA and DIDA

The annotation of human speech in the AIBO corpus considered three individual levels. For the initial phonetic transcription of utterances, the ASCII-based phonetic alphabet SAMPA (Speech Assessment Methods Phonetic Alphabet; http://www.phon.ucl.ac.uk/home/sampa/), which is a subset of the IPA (International Phonetic Alphabet), has proven useful (cf. Leech, 2005). The characters used in the annotation are listed in Appendix A.1 on page 20. Probably, a pseudo-phonetic transcription would have fulfilled the requirements of the small corpus study,

however, in order to maintain the ability of automatically converting the transcription into several other formats, which allows for a long-term use of the annotation, the more formal alternative was favoured. The following example demonstrates the format of the SAMPA:

```
DIs Is @ fo:netik tr@nskrIptSn baI sempa
```

The phonetic transcription was extended by the inclusion of segmental characteristics of utterances (e.g. pauses, utterance breaks or unnoticeable utterances). For this, the DIDA format was chosen. It originates from a discoursive database of the Institute for German Language (Institut für Deutsche Sprache, IDS). For an overview of the transcription format see Appendix A.2. On the basis of the annotation of segmental characteristics of utterances, it could be possible to analyse the relation between the utterance of one participant and the action of another one. Thus, it could be possible to reconstruct the cause of an abruption of an utterance with respect to the non-verbal interaction which is going on.

Another dimension of speech was planned to be annotated: the prosodic features of utterances. For this purpose, the system GToBI (German Tones and Break Indices; cf. Grice & Baumann, 2000) seemed to be suitable. However, the annotation turned out to be too complex and time-consuming. Therefore, though the prosody is an interesting aspect and it would be imaginable to relate it to other modalities within the interaction, it was decided not to annotate the prosody of utterances. Instead, a much simpler part-of-speech annotation was carried out. The underlying annotation scheme is presented in Appendix A.3 on page 21.

3.1.2 Gesture: The Berlin Dictionary of Everyday Gestures

The basis for the annotation of gestures in the AIBO corpus is the Berlin Dictionary of Everyday Gestures (cf. Serenari et al., 2002). The dictionary contains 150 families

of gestures which are provided in the form of video recordings. The descriptions of the individual prototypical gestures served as an orientation point for the creation of an annotation scheme for gestures in the AIBO corpus.

Because of the size and complexity of the Berlin Dictionary, only some interesting entries of gestures were considered, which led to the implementation of six types of everyday gestures and their subtypes in the annotation scheme (cf. Appendix B). The following scene demonstrates the typification of a gesture:

Figure 2: Example of gesture (A10 01 - Holding index finger in a certain direction)

This particular gesture was later annotated by the category „A10 01 - Holding index finger in a certain direction“. Naturally, the annotation scheme can be extended to suit other contexts, but at this point in time it proved satisfactory.

3.1.3 Facial Expression: MPEG-4 Facial Animation Parameters

MPEG-4 SNHC (Moving Pictures Expert Group, Synthetic/Natural Hybrid Coding) seems to be an adequate basis for the annotation of facial expression. The MPEG is a working group of the ISO (International Standardization Organization) and generally deals with the processing of moving pictures in combination with audio data (cf.

Kipp et. al., 2002, p. 23). An outcome of the working group is the coding module MPEG-4, which „enables integration of face animation with multimedia communications and representations and allows face animation over low bit rate communication channels, for point-to-point as well as multi-point connections with low-delay“ (Kipp et al., 2002, p. 24).

At first sight this is a slightly different application than the desired one, namely the animation of facial expression for artificial agents. However, the basic notions of MPEG-4 constitute a useful basis for the annotation of human facial expression. The fundament of MPEG-4 is a model of a prototypical, neutral facial expression imitating a relaxed human face. According to the model, 84 different feature points can be referenced (cf. Kipp et al., 2002, p. 26):

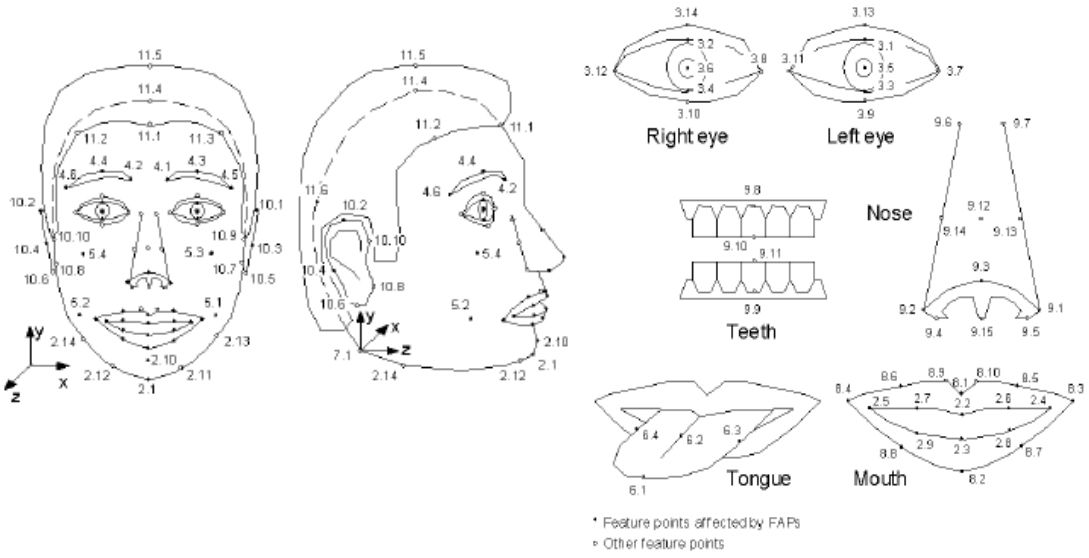


Figure 3: Feature points of MPEG-4

The definition of FAPs, Facial Animation Parameters, and FAPUs, Facial Animation Parameter Units (measurements for the relative distance of individual feature points), allow for the annotation of moving feature points and therefore the annotation of facial expressions. For the AIBO corpus 68 different FAPs were considered (cf. Appendix C, p. 23), whereas the FAPUs were not incorporated because of the

complexity of this task. An example of a facial expression which can be categorised by MPEG-4 is presented in the following figure:

Figure 4: Example of facial expression (35 – raise_l_o_eyebrow + 36 - raise_r_o_eyebrow)

3.1.4 AIBO: LED and acoustic signals

While viewing the recorded sequences, the modalities of AIBO turned out to be a crucial factor for the interaction between AIBO and human beings on the one hand, and the interaction between the human participants on the other hand. The most obvious communicative resources of AIBO are its acoustic signals and the LED display on its head. There are two different acoustic signals: (1) AIBO has a fixed set of sentences and phrases which can be interpreted by the means of human language. These were annotated by the SAMPA which was introduced in section 3.1.1; (2) AIBO is able to produce beep sounds which could be interpreted by the expertise of the AIBO owner, and were annotated accordingly:

Sound	Interpretation
1	Countdown for photo (without taking photo itself)
2	"I just play something right now" or reaction to petting because of touching back sensor while lifting from station
3	Sound corresponding to "Don't drop me, please"
4	Sound corresponding to command "Beg for it"
5	Sound for confirming command "Let's dance"
6	Sound corresponding to AIBOs utterance "Oh yes"
7	Sound corresponding to AIBOs utterance "Here I go"
8	"I am happy about petting"

(Download: http://www.jettka.de/AIBO/AIBO-sounds.zip)

As a second important modality the LED display, which can show several statuses, should have been included in the annotation.

Figure 5: The LED display of AIBO

Unfortunately, the importance of the display for the interaction was not recognised in particular before and during the recording of the data, so that most of the LED statuses are not visible. This circumstance is a clear shortcoming of the recorded data which leads to the premise to draw very special attention to this aspect in later recordings. Despite the lack of visible occurrences, it was decided to implement an annotation scheme, with regard to possible later research (cf. Appendix D, p. 26).

3.2 Annotation tools

Since it does not seem very promising to undertake a manual annotation, the use of some kind of tool is advisable for the annotation of corpora,. There is a wide range of tools available, e.g. ELAN, ANVIL, TASX, the MATE/NITE workbench, Signstream, etc. According to Wittenburg (2008, p. 681), generally several tasks should be fulfilled by annotation tools for multimodal corpora:

- the synchronisation of recordings and annotations
- the definition of annotation tiers and hierarchies between those
- an easy navigation and segmentation
- the query of annotations and patterns (e.g. by regular expressions)
- the export and import of other formats (for usage in other annotation tools)

Since it emerged that ELAN has the complete functionality with regard to these requirements, it was chosen as annotation tool for the AIBO corpus. This does not mean that the other tools are in any way less adequate; there were simply no experiences with them. For a detailed introduction to ELAN see Hulsbosch (2009).

The annotation schemes which were discussed in the previous section mostly could be implemented directly in ELAN by using “Controlled Vocabularies”. These allow for a simple selection of an appropriate category from the annotation scheme (vocabulary) for the annotation of a single segment. The modalities, which should be annotated for the individual participants, could be determined by the means of tiers. Figure 6 shows the tiers of a specific participant. By marking an area on the time axis an annotation segment can be fixed (green selection) and in case a vocabulary has been implemented for the current tier (a modality), a category can be chosen to annotate the segment. If there is no vocabulary, a free annotation can be entered. In the sample below only two tiers allow for free annotations: the commentary tier and the one for the phonetic transcription which is geared to the SAMPA. There is a

simple reason for not implementing a vocabulary for the SAMPA, namely not every single phoneme should be annotated as an individual segment.

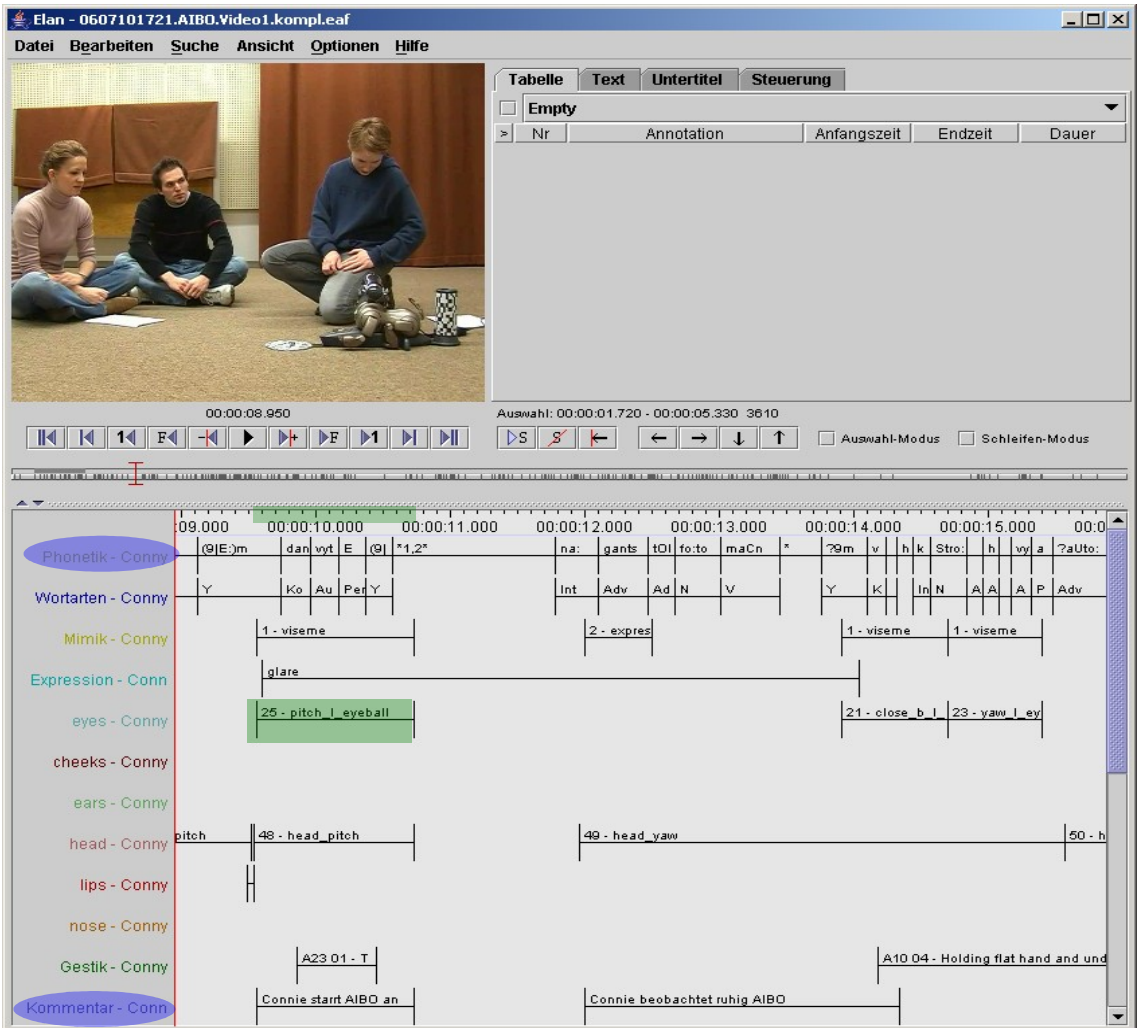


Figure 6: ELAN annotation window

The annotation will not be presented in much detail here. The best overview can be reached by having a look to it in ELAN itself. A sample annotation from the AIBO corpus is available online: http://www.jettka.de/AIBO/ELAN-sample.zip.

The underlying annotation format of ELAN is XML. All user-defined data from ELAN annotations is stored in standoff XML markup. This is a very important feature of the annotation tool since it allows for the flexible and cooperative handling of annotations (cf. Wittenburg, 2008, pp. 677ff).

4 Analysis

Unfortunately, the analysis of the AIBO corpus was not carried out as broadly as desired due to a lack of time in the end of the project. Therefore, just a few basic interpretations of the data, e.g. that the comprehension of AIBO's communication channels is a crucial factor for the interaction, could be made on the basis of the relations between individual modalities (tiers) in ELAN. Another method of analysis within ELAN were the direct search of annotation segments on the basis of their values and the values of parallel annotation segments on other tiers. These methods mainly represent qualitative methods which, for instance, could be used for the generation of hypotheses.

However, quantitative methods are much more in use in corpus linguistics, since their applicability (normally corpora include a large amount of annotated data) and the resulting empirical and statistical conclusions represent a major advantage of corpora over the acceptability judgements of earlier linguistic research (cf. section 2.1). The statistical methods range from relatively simple, but common, frequency tests or co-occurrence statistics to more advanced ones like "the χ^2-test, the *t*-test, the *z*-score, Mutual Information, the binomial test, the Poisson measure, the Fisher-Yates exact test, etc." (Gries, 2006). As the only statistical data which can be derived from corpora is frequency data, Gries (2006) shows that it is eminently important to use more powerful statistical tests to overcome the possibly misleading nature of raw frequency data.

These considerations should be taken into account when it comes to a more complex analysis of a corpus.

5 Conclusion

The goal of this essay is to give an introduction to the basic concepts of multimodal corpora. In order to accompany the underlying theoretical assumptions with some practical experiences, the AIBO corpus was introduced. The corpus focusses on the interaction between human beings and the robot dog AIBO. However, it must be stressed that it was created in the context of a seminar, i.e. it was not primarily meant to be used in advanced academic research, but mainly for educational purposes. Therefore, there are some shortcomings of the AIBO corpus which have to be regarded, e.g. its relatively small size and the lack of a statistical rating of the annotations. Moreover, the outstanding of statistical analyses of the corpus does not allow for any conclusions regarding its appropriateness to capture important details of the interaction between humans and AIBO.

Nevertheless, the basic characteristics and requirements of multimodal corpora were demonstrated by addressing the building (chapter 2), the annotation (chapter 3) and the analysis (chapter 4) of a multimodal corpus. Accordingly, the essay has shown that there are some fundamental steps in the creation of multimodal corpora. In order to reach a high degree of sustainability, the creator should pay special attention to present standards for the format of recordings, metadata and annotations. This does not only facilitate the re-usability and multi-functionality of a corpus, but also allows for the application of various tools for the administration, annotation and analysis which could be problematic for completely new and independent formats.

6 References

Allwood, J., 2008. Multimodal Corpora. In Lüdeling, A. & Kytö, Merja (eds.). *Corpus Linguistics. An International Handbook.* Handbooks of Linguistics and Communication Science. Berlin/New York: de Gruyter, pp. 207-225.

Grice, M. & Baumann, S., 2000. Deutsche Intonation und GToBI. In *Linguistische Berichte*, Vol. 181, Hamburg: Helmut Buske Verlag.

Gries, S., 2006. Some Proposals towards a More Rigorous Corpus Linguistics. In *Zeitschrift für Anglistik und Amerikanistik. A Quarterly of Language, Literature and Culture.* Vol. 54, No. 2, pp. 191-202.

Hulsbosch, M., 2009. *EUDICO Linguistic Annotator (ELAN). Version 3.8. Manual.* [http://www.mpi.nl/corpus/manuals/manual-elan.pdf (09.01.2009)]

Institut für Deutsche Sprache, 2001. *Transkriptionsrichtlinien für die Eingabe in DIDA*. [http://www.ids-mannheim.de/prag/dida/dida-trl.pdf (09.01.2009)]

Kipp, M., Reithinger, N., Bernsen, N., Ole/Dybkjær, L., Wegener Knudsen, M., Machuca, M. & Riera, M., 2002. *Best practice gesture, facial expression, and cross-modality coding schemes for inclusion in the workbench.* [http://nite.nis.sdu.dk/deliverables/NITE-D2.3-F.pdf (09.01.2009)]

Leech, G., 2005. Adding Linguistic Annotation. In Wynne, M. (ed.), *Developing Linguistic Corpora: a Guide to Good Practice.* Oxford: Oxford Books, pp. 17-29. [http://ahds.ac.uk/linguistic-corpora/ (09.01.2009)]

McEnery, T. & Wilson, A., 2008. *Corpus Linguistics. An Introduction.* 2nd edition. Edinburgh: University Press.

Serenari, M., Dybkjær, L., Heid, U., Kipp, M. & Reithinger, N., 2002. *Survey of existing gesture, facial expression, and cross-modality coding schemes.* [http://nite.nis.sdu.dk/deliverables/NITE-D2.1-sept02-F.pdf (09.01.2009)]

Wittenburg, P., 2008. Preprocessing multimodal corpora. In Lüdeling, A. & Kytö, Merja (eds.), *Corpus Linguistics. An International Handbook.* Handbooks of Linguistics and Communication Science. Berlin/New York: de Gruyter, pp. 664-685.

Appendix

A Annotation of speech

A.1 Character set of the SAMPA (German)

Symbol	Word	Annotation
p	*Pein*	paIn
b	*Bein*	baIn
t	*Teich*	taIC
d	*Deich*	daIC
k	*Kunst*	kUnst
g	*Gunst*	gUnst
?	*Verein*	fE6"?aIn
pf	*Pfahl*	pfa:l
ts	*Zahl*	tsa:l
tS	*deutsch*	dOYtS
dZ	*Dschungel*	"dZUN=l
f	*fast*	fast
v	*was*	vas
s	*Tasse*	"tas@
z	*Hase*	"ha:z@
S	*waschen*	"vaS=n
Z	*Genie*	Ze"ni:
C	*sicher*	"zIC6
j	*Jahr*	ja:6
x	*Buch*	bu:x
h	*Hand*	hant
m	*mein*	maIn
n	*nein*	naIn
N	*Ding*	dIN
l	*Leim*	laIm
R	*Reim*	RaIm
I	*Sitz*	zIts

Symbol	Word	Annotation
Y	*hübsch*	hYpS
9	*plötzlich*	"pl9tslIC
i:	*Lied*	li:t
e:	*Beet*	be:t
E:	*spät*	SpE:t
a:	*Tat*	ta:t
o:	*rot*	ro:t
u:	*Blut*	blu:t
y:	*süß*	zy:s
2:	*blöd*	bl2:t
aI	*Eis*	aIs
aU	*Haus*	haUs
OY	*Kreuz*	krOYts
@	*bitte*	"bIt@
6	*besser*	"bEs6
i:6	*Tier*	ti:6
I6	*Wirt*	vI6t
y:6	*Tür*	ty:6
Y6	*Türke*	"tY6k@
e:6	*schwer*	Sve:6
E6	*Berg*	bE6k
E:6	*Bär*	bE:6
2:6	*Föhr*	f2:6
96	*Wörter*	"v96t6
a:6	*Haar*	ha:6
a6	*hart*	ha6t
u:6	*Kur*	ku:6

Symbol	Word	Annotation
E	*Gesetz*	g@"zEts
a	*Satz*	zats
O	*Trotz*	trOts
U	*Schutz*	SUts

Symbol	Word	Annotation
U6	*kurz*	kU6ts
o:6	*Ohr*	o:6
O6	*dort*	dO6t

:	*noticeable lengthening*
"	*noticeable stress*

::	*very long lengthening*

A.2 Transcription of suprasegmental characteristics: DIDA

Annotation	Meaning
/	*word abruption*
*****	*short pause of speech*
******	*long pause of speech*
1,50	*Pause of one and a half seconds*
3:30	*Pause of three minutes and thirty seconds*
+	*direct attachment of an utterance by one to that by another speaker*
(this may be)	*assumed wording*
(day?bay)	*alternative wording*
(...)	*unnoticeable utterance; if possible 3 dots per syllable*
(...2,5)	*length of unnoticeable utterance*
[...]	*omission in transcript*
[10min]	*omission in transcript with absolute length*

A.3 Part-of-speech tag set

POS	Annotation
Abbreviation	**Abk**
Adjective	**A**
Adverb	**Adv**
Determiner	**Art**

POS	Annotation
Noun	**N**
Participle1	**P1**
Participle 2	**P2**
Personal Pronoun	**PerPro**

POS	Annotation
Auxiliary Verb	**Aux**
Demonstrative Pronoun	**DemPro**
Proper Name	**Eig**
Indefinite Pronoun	**IndPro**
Interjection	**Int**
Interrogative Pronoun	**IntPro**
Conjunction	**Konj**
Modal Verb	**Mod**

POS	Annotation
Possessive Pronoun	**PosPro**
Preposition	**Präp**
Reflexive Pronoun	**RefPro**
Relative Pronoun	**RelPro**
Verb	**V**
Verb Particle	**VPar**
no POS found	**Y**
Numeral	**Z**

B Extract of the Berlin Dictionary of Everyday Gestures

A01 ___	**Wafting a hand in front of one's eyes**
A01 01	Sliding flat hand to the side in front of face
A01 02	Sliding flat hand to the side in front of face and looking to the floor
A01 03	Waving splayed hand to and fro in front of face
A01 04	Waving splayed hand to and fro in front of face with head pushed forwards

A10 &C04-E03 ___	**Pointing**
A10 &C04-E03	<01><02><03><04><05><06><07><08><09> Pointing with hand
A10 &C04-E03 01	Holding index finger in a certain direction
A10 &C04-E03 02	Holding index finger forwards with touch side directed downwards
A10 &C04-E03 03	Holding index finger forwards with touch side directed upwards
A10 &C04-E03 04	Holding flat hand and underarm in a particular direction
A10 &C04-E03 05	Holding flat hand with touch side directed downwards in a particular direction
A10 &C04-E03 06	Holding pointed out thumb in a particular direction
A10 &C04-E03 07	Directing index finger to chest
A10 &C04-E03 08	Holding index finger alternately in two different directions
A10 &C04-E03 09	Pointing index finger to addressee and to chest
A10 &C04-E03	<10><11><12> Pointing with other parts of body
A10 &C04-E03 10	Pushing chin forwards
A10 &C04-E03 11	Tilting head
A10 &C04-E03 12	Rolling eyes in a particular direction

A13 ___ **Lowering flat hands**

A13 01 Lowering hand and arm at stomach height

A13 02 Lowering finger of flat hand to stomach height

A13 03 Pushing vertical flat hand at stomach height forwards and downwards

A19 ___ **Twiddling one's thumbs**

A19 01 Twiddling thumbs around each other

A19 02 Twiddling thumbs around each other quickly

A19 03 Turning thumbs forwards and backwards around each other

A23 ___ **Twisting one's hand back and forth**

A23 01 Twisting horizontally splayed out hand

A23 02 Turning splayed out hand to and fro and displaying the side

A23 03 Repeatedly turning away vertical, frontally displayed, splayed out hand

A26 ___ **Raising flat hands**

A26 01 Raising horizontal flat hand

A26 02 Raising horizontal flat hand quickly and repeatedly

A26 03 Tipping flat hand upwards

A26 04 Raising flat hand horizontally after circling movement

A26 05 Tipping flat hand several times upwards and continuing to raising it

(cf. Serenari et al., 2002, pp. 36ff)

C Extract of MPEG-4 Facial Animation Parameters

#	FAP name	FAP description	Units	Uni- or bidirectional	Position motion	Group	FDP subgroupnumber
1	viseme	Set of values determining the mixture of two visemes for this frame (e.g. pbm, fv, th)	na	na	na	1	na
2	expression	A set of values determining the mixture of two facial expression	na	na	na	1	na
3	open_jaw	Vertical jaw displacement (does not affect mouth opening)	MNS	U	down	2	1
4	lower_t_midlip	Vertical top middle inner lip displacement	MNS	B	down	2	2

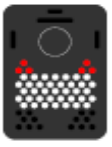
5	raise_b_midlip	Vertical bottom middle inner lip displacement	MNS	B	up	2	3
6	stretch_l_cornerlip	Horizontal displacement of left inner lip corner	MW	B	left	2	4
7	stretch_r_cornerlip	Horizontal displacement of right inner lip corner	MW	B	right	2	5
8	lower_t_lip_lm	Vertical displacement of midpoint between left corner and middle of top inner lip	MNS	B	down	2	6
9	lower_t_lip_rm	Vertical displacement of midpoint between right corner and middle of top inner lip	MNS	B	down	2	7
10	raise_b_lip_lm	Vertical displacement of midpoint between left corner and middle of bottom inner lip	MNS	B	up	2	8
11	raise_b_lip_rm	Vertical displacement of midpoint between right corner and middle of bottom inner lip	MNS	B	up	2	9
12	raise_l_cornerlip	Vertical displacement of left inner lip corner	MNS	B	up	2	4
13	raise_r_cornerlip	Vertical displacement of right inner lip corner	MNS	B	up	2	5
14	thrust_jaw	Depth displacement of jaw	MNS	U	forward	2	1
15	shift_jaw	Side to side displacement of jaw	MW	B	right	2	1
16	push_b_lip	Depth displacement of bottom middle lip	MNS	B	forward	2	3
17	push_t_lip	Depth displacement of top middle lip	MNS	B	forward	2	2
18	depress_chin	Upward and compressing movement of the chin (like in sadness)	MNS	B	up	2	10
19	close_t_l_eyelid	Vertical displacement of top left eyelid	IRISD	B	down	3	1
20	close_t_r_eyelid	Vertical displacement of top right eyelid	IRISD	B	down	3	2
21	close_b_l_eyelid	Vertical displacement of bottom left eyelid	IRISD	B	up	3	3
22	close_b_r_eyelid	Vertical displacement of bottom right eyelid	IRISD	B	up	3	4
23	yaw_l_eyeball	Horizontal orientation of left eyeball	AU	B	left	3	5
24	yaw_r_eyeball	Horizontal orientation of right eyeball	AU	B	left	3	6
25	pitch_l_eyeball	Vertical orientation of left eyeball	AU	B	down	3	5
26	pitch_r_eyeball	Vertical orientation of right eyeball	AU	B	down	3	6
27	thrust_l_eyeball	Depth displacement of left eyeball	ES	B	forward	3	5
28	thrust_r_eyeball	Depth displacement of right eyeball	ES	B	forward	3	6
29	dilate_l_pupil	Dilation of left pupil	IRISD	B	growing	3	5
30	dilate_r_pupil	Dilation of right pupil	IRISD	B	growing	3	6
31	raise_l_i_eyebrow	Vertical displacement of left inner eyebrow	ENS	B	up	4	1
32	raise_r_i_eyebrow	Vertical displacement of right inner eyebrow	ENS	B	up	4	2
	FAP name	**FAP description**	Units	Uni- or bidirectional	Position motion	Group	FDP subgroupnumber
33	raise_l_m_eyebrow	Vertical displacement of left middle eyebrow	ENS	B	up	4	3
34	raise_r_m_eyebrow	Vertical displacement of right middle eyebrow	ENS	B	up	4	4
35	raise_l_o_eyebrow	Vertical displacement of left outer eyebrow	ENS	B	up	4	5
36	raise_r_o_eyebrow	Vertical displacement of right outer eyebrow	ENS	B	up	4	6

37	squeeze_l_eyebrow	Horizontal displacement of left eyebrow	ES	B	right	4	1
38	squeeze_r_eyebrow	Horizontal displacement of right eyebrow	ES	B	left	4	2
39	puff_l_cheek	Horizontal displacement of left cheek	ES	B	left	5	1
40	puff_r_cheek	Horizontal displacement of right cheek	ES	B	right	5	2
41	lift_l_cheek	Vertical displacement of left cheek	ENS	U	up	5	3
42	lift_r_cheek	Vertical displacement of right cheek	ENS	U	up	5	4
43	shift_tongue_tip	Horizontal displacement of tongue tip	MW	B	right	6	1
44	raise_tongue_tip	Vertical displacement of tongue tip	MNS	B	up	6	1
45	thrust_tongue_tip	Depth displacement of tongue tip	MW	B	forward	6	1
46	raise_tongue	Vertical displacement of tongue	MNS	B	up	6	2
47	tongue_roll	Rolling of the tongue into U shape	AU	U	concave upward	6	3, 4
48	head_pitch	Head pitch angle from top of spine	AU	B	down	7	1
49	head_yaw	Head yaw angle from top of spine	AU	B	left	7	1
50	head_roll	Head roll angle from top of spine	AU	B	right	7	1
51	lower_t_midlip _o	Vertical top middle outer lip displacement	MNS	B	down	8	1
52	raise_b_midlip_o	Vertical bottom middle outer lip displacement	MNS	B	up	8	2
53	stretch_l_cornerlip_o	Horizontal displacement of left outer lip corner	MW	B	left	8	3
54	stretch_r_cornerlip_o	Horizontal displacement of right outer lip corner	MW	B	right	8	4
55	lower_t_lip_lm _o	Vertical displacement of midpoint between left corner and middle of top outer lip	MNS	B	down	8	5
56	lower_t_lip_rm _o	Vertical displacement of midpoint between right corner and middle of top outer lip	MNS	B	down	8	6
57	raise_b_lip_lm_o	Vertical displacement of midpoint between left corner and middle of bottom outer lip	MNS	B	up	8	7
58	raise_b_lip_rm_o	Vertical displacement of midpoint between right corner and middle of bottom outer lip	MNS	B	up	8	8
59	raise_l_cornerlip_o	Vertical displacement of left outer lip corner	MNS	B	up	8	3
60	raise_r_cornerlip _o	Vertical displacement of right outer lip corner	MNS	B	up	8	4
61	stretch_l_nose	Horizontal displacement of left side of nose	ENS	B	left	9	1
62	stretch_r_nose	Horizontal displacement of right side of nose	ENS	B	right	9	2
63	raise_nose	Vertical displacement of nose tip	ENS	B	up	9	3
64	bend_nose	Horizontal displacement of nose tip	ENS	B	right	9	3
65	raise_l_ear	Vertical displacement of left ear	ENS	B	up	10	1
66	raise_r_ear	Vertical displacement of right ear	ENS	B	up	10	2
67	pull_l_ear	Horizontal displacement of left ear	ENS	B	left	10	3
68	pull_r_ear	Horizontal displacement of right ear	ENS	B	right	10	4

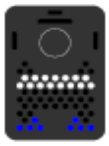
(cf. Kipp et al., 2002, pp. 28ff)

D The LED display of AIBO

AIBOs basic "mental" states		
LED display	Description	Annotation
	happy	**glückl.**
	bad mood	**schl. gel.**
	sad	**traurig**
	angry	**böse**
	afraid	**ängst.**
	surprised	**überr.**

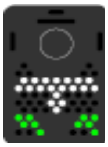
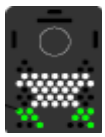
AIBOs basic messages		
LED display	Description	Annotation
	discovered AIBONE	**AIBONE erk.**
	discovered pink ball	**Ball erk.**
	discovered human face	**Gesicht erk.**
	discovered mistress/lording	**Herrchen erk.**
	infra-red sensor (for distances) in head has discovered barrier	**Kopf Hind.**
	infra-red sensor (for distances) in chest has discovered barrier	**Brust Hind.**

status		
LED status	Description	Annotation
	illuminated blue: AEP switched on (autonomous mode)	**AEP autonom**
	blinking blue: AEP switched on (remote mode)	**AEP ferng.**
	illuminated pink: play mode	**Spielmodus**
	illuminated green: AIBO is relaxing	**ruht aus**
	blinking green slowly: AIBO wants to be lifted	**will hoch**
	illuminated or blinking yellow **clinical mode**	**Klinik**

	legs on right hand side have discovered barrier (crash)	**re. Beine Hind.**
	legs on left hand side have discovered barrier (crash)	**li. Beine Hind.**
	head has discovered barrier (crash)	**Kopf Ges. Hind.**
	AIBO discovered human hand	**Pfötchen Hand erk.**
	distance sensors in chest discovered falling item diagonal in front	**Brust fall. Geg.**
	pink item discovered	**rosa Geg.**

favourite situations		
LED status	Description	Annotation

	illuminated red: alarm mode (high temperature)	**Alarm**

	AIBO inspects his favourite place	**Liebl.platz**
	AIBO discovered one of favourite items	**Liebl.geg.**
	AIBO being touched	**Anfassen**

(cf. AIBO ERS-7 – http://www.sonydigital-link.com/aibo/downloads/de/ledface.pdf)

E AIBOs speech commands

Locomotion

"Stand up"
"Sit down"
"Lay down"
"Go forward"
"Go back"
"Turn left"
"Turn right"
"Turn around"
"Come here"
"AIBO come here"
"come here AIBO"
"Over here"

Interaction with toy

"Find your AIBONE"
"Pick up your AIBONE"
"Bring me your AIBONE"
"Give it to me"
"Open your mouth"
"Find your ball"
"Kick the ball"

Settings

"Name registration"
"Owner registration"

Simple actions

"Bark"
"Bark Bark"
"Let's Dance"
"Shake hands"
"The other paw"
"Beg for it"
"Chase your tail"
"Take a picture"
"Snapshot"
"Look around"
"Seven Strike!"
"Don't do that"

"Walk around"

"Go to the station"

"Stop"

"Wait"

"Go"

"Move it"

Questioning

"What's up"

"How are you?"

"What's going on?"

"What are you doing?"

"What time is it?"

"Are you sleepy?"

"Are you tired?"

"Are you hungry?"

"What's your name?"

"What's your owner's name?"

"Do you love me?"

"Where is your ball"

"Where is your pink ball"

"Where is your AIBOne"

"Where is your room"

"Where is your station"

"I love you AIBO"

"I am your owner"

"Set alarm"

"Favorite thing registration"

Conversation

"Snowy"

"AIBO"

"yes"

"no"

"Hello"

"Good morning"

"Good Evening"

"I'm here"

"Say hello"

"Good Bye"

"Bye Bye"

"Good night"

"I'm leaving"

"I'm leaving now"

"See you later"

"Good boy"

"Good AIBO"

"Good girl"

"How cute"

"Go for it"

"school"

"nowhere"

"Be careful"

"I have good news"

"I had a bad day"

"I'm sleepy"

"I'm tired"

"I'm fine"

"I'm okay"

"I'm good"

"I'm starving"

"I'm hungry"

"It was fine"

"It was nice"

"It was okay"

"It was good"

"it was boring"

"Be Quiet"

"just kidding"

"I'm sorry AIBO"

"I'm sorry"

"AIBO I'm sorry"

"work"