Managing the Speech flow

—Towards a working definition of utterance for use in CHAT-coded transcripts—

Susanne Miyata

1. Problems with the practical definition of utterance

It is usual to define early child language production as an 'utterance' rather than 'sentence', because, as Bloom (1973: 55) suggested, children at the very beginning of language acquisition do "not as yet know the linguistic code for mapping conceptual notion onto semantic-syntactic relations in sentences". In this sense everything the child utters in a row (or in other words everything belonging to a coherent intonation contour) forms one utterance, even if it is doubtful whether the elements are connected grammatically. For early child language productions it is convenient to use the term 'utterance' rather than 'sentence', because it is avoids a judgement about the grammatical status of the production. Since the decision of what constitutes an utterance, semantic and paralinguistic, and also the immediate context, as well as in those fortu-

nate cases the historical context (that is previous language production and events) add to the judgement of what is an utterance. For example, Blake / Quartaro / Onorati (1993: 142) consider cues as "long pauses, intonation, intervening turns by the experimenter, and the presence or absence of connectives" for the judgment of utterance boundaries of the speech productions of their 1;6–4;9 year old children for the purpose of MLU count.

Nevertheless the term 'utterance' is used for non-infant speakers as well, and seems to have there another nuance. The problem with adult speech productions lies not in the decision, whether something is already a sentence or not yet, but rather whether elliptic responses, ill-formed sentences as well as formulaic expressions can be called sentential. The 'utterance' concept tries to avoid this decision, by laying the stress on the intonational coherence. This is certainly due to influence from interactional research, which rather relies on units like speaker turn and intonation unit (cf. DuBois / Schuetze-Coburn 1993) to analyze the speech flow.

The concept of utterance though, seems to rely basically on the sentence, while allowing a more generous interpretation of the structure by adding intonational cues. So "ill-formed sentences" can be included without falling into the logical trap (if a sentence is defined as a "well-formed structure", as Trask 1993 concludes). Also any items which stand outside the syntactical structure like 'ehm', laughing or a gesture, can be integrated. It seems to be just the mixture of grammatical and interactional criteria for the judgement of utterance boundaries which causes problems for the utterance definition, in a practical as well as theoretical way.

The two notions of utterance and sentence exemplify two divergent theoretical positions, as utterance emphasizes the actual speech which occurs in a historical context from a more or less interactional point of view, while for sentence the focus of interest lies not so much in the actual speech productions but in the underlying ideal structure, independent of historical production.

It is doubtful whether this bisection is fruitful in the long run. It is possible, desirable, and inspiring to examine grammatical structures not only in ideally constructed structures, but also in real historical sentences. Confronted with historical sentences it becomes necessary to deal with data contradictory to theory. This can give a new impulse for theoretical thinking about grammar.Dealing with real historical data, it becomes a challenge to reconstruct the internal representation of grammatical structure for the individual speaker, child or adult, learner of his first language or a second. One can assume that a non-infant language learner applies the notion of 'sentence' to any speech production in the second language as well, not matter how far from the

target language's ideal structure this speech string may be. It is therefore a task to analyse this individual "interlanguage" (Selinker) in order to understand the processes underlying language acquisition.

The use of actual speech data for grammatical analysis has become not only desirable but also practically possible by the use of publically available transcriptions of natural speech data as collected by CHILDES. The Child Language Exchange System CHILDES (MacWhinney 1995) has developed a transcription format named CHAT which provides a refined set of rules for transcription of natural speech. For CHAT (and with it the CLAN analysis programs) the basic analysis items are words and utterances. In addition to these two levels, it is possible to focus on the phonetic level (coding in IPA or UNIBET), on morpheme level (separating prefixes and suffixes, doing automatic morpheme analysis with MOR or MLU counting), and on speaker turns (treating all utterances of a turn as one unit). Nevertheless as utterances are the basic item and starting point for many analyses, interactional as well as grammatical, the decision of what constitutes an utterance strongly influences the analysis results. In spite of its crucial importance, no definition of utterance are provided, leaving the definition completely up to the researcher. This omission is understandable, as CHILDES intends itself to be an objective tool for language research, intending to remain neutral within any theoretical frame.

As an example Terada 1994: 170 indicates in her critical review of possibilities of transcription of L2 data within CHILDES, in the following interaction between a L2 Japanese learner and a native Japanese teacher, it is left to the subjective judgement of the transcriber whether the learner utters 1, 2, or 3 utterances (Terada 1994: 170).

(1) Learner:	aa	kono	e	wa	soo	inu	desu	ne ?
	oh	this	picture	TOPIC	right	dog	COPULA	TAGQ
Teacher:				un				
				yeah				

Terada presents the following four possibilities for transcription of the learner's utterance:

(a) aa kono e wa.... [trailing off]
soo inu desu ne.
(b) aa kono e wa soo inu desu ne.
(c) aa kono e wa inu desu ne.
soo. [interpolated indepedent utterance]
(d) aa kono e wa.... [trailing off]

soo.
inu desu ne.

Moreover, it could even be possible to count 4 utterances by separating the initial "aa". Clearly, the different solutions are derived from the application of different criteria.
(a) is completely based on intonational cues (following the criterion of the coherent intonation unit), while (c) follows only the grammatical structure. (b) and (d) are mixtures of both criteria. In other words, the mixture of grammatical and interactional criteria for the judgement of utterance boundaries causes the judgement to be ambiguous.

This kind of problem is no longer confined to theory, but has acute relevance to the actual language trancscription for database use. For universally shared data the application of a uniform transcription system is crucial for the usability of the data and the reliability of any analysis results. The definition of utterance as a basic unit influences the outcome of many grammatical analyses.

It is possible that the difficulty of a definition of utterance is felt more strongly in languages which do not focus on the sentence as basic unit. Tsao (1977, cited by Huang 1984) differentiates within the frame of the 0-argument discussion, between sentence oriented languages like English, and discourse oriented languages like Chinese.

In sentence-oriented languages a syntactically complete sentence is requested, even if the arguments are clear from the context, it is necessary to use dummy forms (say pronouns) for omitted subjects or objects. Compare the following English example to their Japanese equivalents.

(2) I'll give it to you.
(3) ageru.
give
"I'll give it to you."
(4) It is dark.
(5) kurai.
is dark
"It is dark."

Discourse-oriented languages on the other hand allow highly elliptic sentences. For example, in Japanese not only the subject and / or the object can be left out, but also trailing off midway is often used stylistically. The following interaction, a well-known

stumbling block for learners of Japanese, may serve as an example.

(6) Client: kinenkitte kudasai.
memorial stamps please give me
Post-office clerk: kinenkitte wa ima chotto...
memorial stamps TOPIC now a bit

The post-office clerk is trailing off before stating that the stamps are actually sold out. In fact what is expected in this case is the reaction of the hearer, be it "a arimasen ka" [there aren't any ?] or "a soo desu ka" [oh I see], which is automatically followed by a "hai" [yes], which will close this unit. In other words, the elliptic answer, together with the reaction of the hearer and the approval of the speaker, will construct one unit.

Another example are the feedback signs (jp.: aizuchi), which are constantly expected from the hearer.

(7) sore de ne kiitemitara moo nai to iwarete kekkyoku sono mama kaetchatta kedo...
and then when I asked they said there aren't anymore so I had to go home without them
un un un ara
yeah yeah yeah oh dear

These feedback signs are not so much independent utterances of the listener, nor do they prepare a turn change, but are rather expected and prepared for by the speaker himself, who signals by intonation and gaze that he is expecting a feedback sign, and waits for it.This can even be observed in speakers who are supplying feedback signs by themselves (for ex. in interview situations).

But, what is important for the argument here, although these feedback signs are part of the speaker's turn, they are not syntactically integrated in his sentence. However the supplying of aizuchi rather strongly influences the view of language of the Japanese speaker, leading it to phrasal (jp.: bunsetsu) rather than sentential units.

As I have mentioned before, in the case of non-infant speakers' data, the actual decision of what to call an utterance relies to a great part on how we conceive the sentence. For sentence-orientated languages the utterance concept gives more flexibility by allowing the inclusion of other elements occuring in natural speech, including intonational cues. For discourse-orientated languages on the other hand, the application of the concept of 'sentence' is not as obvious when dealing with natural language data, but of course that does not mean that discourse-oriented languages cannot be analyzed

on a syntactical level. There are still sentences, elliptical or constructed by speaker and listener together, which can be analyzed as such, although there is a greater portion of items, which are not bound syntactically, and which belong to a different level of speech production.

The claim here is that the reported difficulties of definition of utterance result from mixing these two levels of language production, the syntactic level (somehow grammatically structured / connected strings, which I will call syntagma in the following) and the non-syntactic level (syntactically irrelevant strings, like feedback signs, which will be called non-syntactic elements).

2. The utterance: syntactic and non-syntactic elements

Below I will try to develop an operational coding system which allows syntactical analysis as well as interactional analysis in an automated way, by keeping both levels distinct. This coding system uses the symbols and transcription conventions of CHAT, but can be applied for any speech transcription system.

The utterance in the sense I will use it from here on, consists of a syntagma and non-syntactic elements, and also includes proto-syntactic productions of the early child language. An utterance may contain one and only one syntagma. Utterances without a syntagma (for example feedback signals or "yes" responses) are called 0-syntagma. An utterance may contain an unlimited number of non-syntactic elements.The one- and two-word utterances of infant language for which it is problematic to ask for the syntactic status of the utterance, may remain unanalyzed as proto-syntactical on the utterance level, without classifying them any further as syntagma or non-syntactic.

3. The syntagma: structured strings

A syntagma is a syntactically structured string in the broadest sense. This notion is based on the concept of macrosyntagma, as it was proposed by Loman / Joergensen (1971). The macrosyntagma is defined as "a grammatical cohesive unit which is not part of any larger grammatical construction. Other than written sentences unit in writing, it may vary greatly in length, from a monosyllabic interjection, to a multiword sentence expanded by a large number of subordinate clauses." (after Edwards 1993: 21)

In this definition there are two points to underline. The macrosyntagma is defined more broadly than a sentence, as it includes ellipses and interjections, as well as loosely connected clauses. The other point concerns the upper boundary. The macrosyntag-

ma is by definition "not part of any larger grammatical construction". So in this sense, any items, no matter how loosely they may be connected, are part of the same macrosyntagma (Note 1).

The other point concerns the lower boundary of the definition. Ellipses and interjections which do not have clausal status, are seen as macrosyntagma as well. This allows the inclusion of elliptic responses like 'me' to the question 'who likes icecream ?'. Nevertheless, the inclusion of interjections is contradictory, as interjections are not connected to the grammatical structure.

So in the present study I will define the syntagma as a grammatically cohesive unit which is more or less strongly connected by syntactical and / or morphological devices, and which is not part of any larger grammatical construction. It includes multiclaused strings, as well as minimal elliptic responses, as long as they have syntactical potential, as well as ill-formed and un-completed sentences. In other words, the concept of syntagma presented here is broader than 'sentence', because it allows the inclusion of response ellipses, as well as long stretches of loosely connected subclauses as one big unit, ——even when it is interrupted by other items, longer pauses or turn changes. On the other hand this concept of syntagma is more precise than 'macrosyntagma', because it suppresses any non-syntactical elements occuring during the production of a syntagama, as the overall criterion for belonging to a syntagma is the grammatical cohesiveness.

4. Non-syntactic elements: fillers and feedbacks

The speech flow not only contains syntagma of various lengths, but also non-syntactic elements, which fulfill important communicative functions, while being structurally independent from the syntagma. Non-syntactic elements can also constitute an independent utterance (0-syntagma), when they are used alone or in combination with other non-syntactic elements.

The group of non-syntactic elements contains

a) interactional elements like fillers, also including calling expressions, and paralinguistic elements, like gestures, facial expression, laughing and weeping, as long as they have a communicative function within the interaction, and
b) sociocentric formulas, like greetings.

a) The anarchic (Note 2) group of interactional words and sounds can be divided into items which are part of one's own speech (speaker-inserted), and those which are inserted by the auditor into the speech flow of the speaker (auditor-inserted). Or, in

other words, any non-syntactic item the speaker utters during his turn, is speaker-inserted, while anything the auditor utters at the same time is auditor-inserted.

The speaker-inserted items can be divided into speaker continuation signals, elliptical units which don't have syntactic status, calling expressions, and paralinguistic elements.

Speaker continuation signals (a term proposed by Duncan / Fiske 1985, but used here in a slightly different way) are fillers, which are produced by the speaker,to establish or maintain the turn. They can be vocalizations like "ehem" or "fuun", lexical words like "well", "anoo", "nanka", as well as longer strings like "you see", "or something", "to iu ka".

Elliptical units which don't have a clausal status, often occur in responses like "yes", "hai", where they can accompany or even replace a sentence. However elliptical units, which have a syntactic connection to the question before, do not belong to this group (but are counted, rather, as syntagma). So the answer "yes" to the question "do you like ice cream ?" will be analyzed as a speaker inserted non-syntactic element, while the answer "me" to the question" who likes ice cream ?" will be an ellipitic syntagma.

Calling the conversation partner by his name, his title, or with a pronoun, is another special case of speaker-inserted interactional items.

(8) I have to talk to you, Mary.
(9) kore nani, anta ?
this what you
"you, what's this !?"
(10) sensee, chotto ii desu ka ?
"teacher a bit good COP QPART
"teacher, do you have a second ?"

Any of these expressions can also be used as a part of the syntagma, and sometimes it is not easy to decide what is intended. Especially in null-argument languages like Japanese, where it is usual to drop the subject or object when it is obvious from the context, the status of names or pronouns of the second person may be ambiguous, although intonation can give clues. In the example below, the continuation on the same pitch indicates a syntactic connection in the first case, while in the second the pitch is lowered for the second mora [syllable].

(11) anata nani shita no ?

you what did QPART
"what did you do ?"

(12) anata, nani shita no ?
"you, what did you do ?"

Paralinguistic items like pointing or nodding may also accompany or replace a syntagma, and can be part of the speaker-inserted non-syntactical items, as long as they have communicative intention (Note 3).

The auditor-inserted items or auditor backchannel signals (Note 4), are non-syntactic items which signal the attentiveness of the auditor and supporting the speaker in maintaining his turn. To these auditor backchannel signals belong the above mentioned aizuchi (feedback signals) as well paralinguistic signals like nodding, head shaking, or smiling or grimacing. The auditor backchannel signals are inserted by the auditor into the speech flow of the speaker, and expressing the active participation of the auditor. They can also appear as strings like "I see" or "soo desu ka" [is it so ?]. As they are often expected by the speaker at certain points of his speech flow, as we have seen for the aizuchi before, they can be influenced by, but nevertheless not be part of the syntactic structure, the speaker is developing.

b) The second group of non-syntactical elements contains sociocentric formulas (Note 5) like greetings and other formulas used in social interactions. One characteristic of a formula is its morphological inflexibility. For example it is not possible to construct a plural form "good mornings" or change the tense of the following greeting without losing the greeting character.

(13) odekake desu ka ?
Going-out COP QPART
"Are you going out ?"

(14) *odekake deshoo ka ?
Going-out COP [Future/Poss] QPART
"Will you be going out ?" or alternatively
"Are you possibily going out ?"

(15) *odekake deshita ka ?
Going-out COP [Past] QPART
"Have you been out ?"

However in Japanese with its rich lexicon of formulaic expressions, many formulas vary in politeness and expliciteness. Coulmas / Marui / Reinelt (1983: 157) define formulas as "expressions, which, occuring with high frequency in standardized com-

municative situations, take over specific forms of the interaction of the conversation partners" (Note 6).

We define sociocentric formulas here as high-frequency standardized expressions, which often occur at the beginning and the end of conversations, aiming at the (re)establishment and ratification of the social quality of the relationship of the conversation partners. In this sense an exchange like "ogenki desu ka ? -okagesama de." [How are you ? -Fine, thank you] will be analyzed as sociocentric formula as well as "gambatte kudasai" [Go for it !].

5. Non-interactional elements: communicative elements outside of the utterance

In the interactional flow of a conversation we also find "noisy" elements, like coughing, hiccups, a foreign accent, body movements, or swinging earrings etc. These elements are communicative in the overall situation. For example the color of the shirt can signal a certain political attitude. Nevertheless, unless these elements are thematized, they are not part of the ongoing interaction, and stand outside of the utterance.

Combining the concepts explained above we get the scheme presented in table 1. A grammatical analysis will focus on the syntagma, while interactional analysis will include the non-syntactic elements as well. A powerful transcription system should allow both kinds of analysis, the grammatical analysis as well as the interactional analysis. This can be achieved by transcribing syntactic structures and interactional elements as well as sociocentric formulas on different levels. In the next step we will see how these levels can be presented within the technical frame of CHAT.

6. CHAT coding possibilities

The graphical representation of the different levels on separate lines is not possible in a vertical transcription format, where (not only typographically) everything in an utterance is transcribed on the same line, while every other utterance occupies another line of its own (Note 7).

Nevertheless CHAT offers an elaborate set of symbols for coding grammatical as well as interactional phenomena. Hence morphological structure can be expressed by a # (prefix) or- (suffix) as well as other symbols, intonational features by -? (rising intonation)-. (falling intonation). Interruptions and trailing off, uptakes, retraces, overlaps, and errors can be indicated by symbols like +/. (self-interruption), +... (trailing off), ++ (other-completion), [>] (overlap follows), [/] (retracement without correction),

[*] (error) and groups of words like babytalk expressions can be marked by @ as special words, to mention only some of the coding possibilities (cf. MacWhinney 1995 [online version 1996/8]).

The coding of interruption and uptake gives us the opportunity to analyze interrupted utterances as one unit, and the retracement symbols allow us to ignore false starts or self-corrections, if desired, as well as to thematize them, according to the purpose of research. Especially interesting for our purposes here are the @ special word markers, which are subdivided by adding further letters like @o for onomatopeias. By using the CLAN analysis programs it is possible to analyze a text excluding words ending with @text, or on the contrary focus on them.

7. Keeping syntactic and non-syntactic elements distinct

a) On the main line (utterance line) only syntagma and non-syntactic elements are transcribed. Additionally the dependent tier %gpx (gestures and proxemics) may be used for non-verbal communicative elements. Non-interactional elements (if transcribed at all) are noted on other dependent tiers or header tiers (for example %com or @Situation).

b) One utterance can only contain one syntagma. Each item belonging to a syntagma is part of the same utterance. Inserted independent syntagma are transcribed as separate utterances. Interrupted syntagma (marked by +/. or +/?) can be taken up on a new utterance line using the +, (uptake) symbol. The completion of the syntagma can be done by the speaker as well as by the auditor (in which case the ++ symbol is used). This case is not rare, especially in L2 situations, where the auditor will help the speaker to complete a syntagma. In this case the speaker 'owns' the structure as far as he produced it,while the auditor 'owns' the whole structure making the utterance of the speaker his own.

c) The elements on the main line are marked by @is, @ic, @ia, or @if, if they are non-syntactic elements, and unmarked, if they belong to a syntagma.

d) The non-syntactic elements can be divided into the following 5 groups.

un@ia	interactional elements: auditor-inserted
anoo@is	interactional elements: speaker-inserted
anata@isc	interactional elements: speaker-inserted: calling (subgroup of @is)
odekake+desu+ka@if	sociocentric formula

The separation of @isc from @is is due to special status of calling expressions within the speaker-inserted elements.

e) Sociocentric formulas containing two or more words can be connected by a + symbol to treat them as one unit.

kono+aida+wa+doomo+arigatoo+gozaimashita@if
un+un+un@ia

f) paralinguistic elements can be transcribed using the [=! text] symbol or the %gpx tier. In order to allow automated analysis the use of standardized expressions like "pointing", "laughing" or "yubisashi", "warai" is helpful. If they are accompanied by vocalizations, these can be transcribed as "non-words" using the & symbol.

soo@ia [=! warai].
&hehe [=! warai].

g) non-interactional elements can be transcribed on %act or other tiers or as vocalizations using the & symbol, which assigns them a status as non-word.

&hakSoN [%com: sneezing].

Within CHILDES, this simple coding system allows the systematic ex- or inclusion of non-syntactic elements to be automatically conducted by using the -s*@i* (for exclusion) or +s*@i* option (for inclusion), depending on the goal of the analysis. So it is possible to focus on aizuchi (feedback signals) by looking up all items marked with @ia, while being possible to ignore the aizuchi when concentrating on the syntactic structure.

Furthermore, for grammatical analysis (for example when using MOR or MLU) the non-syntactic elements will be ignored, and also feedback signals will constitute 0-syntagma, and whatever the decision of the transcriber may be ——e.g. whether to afix a non-syntactic element to the completed syntagma or the following, or on the contrary to give it the status of an independent utterance, ——it will not influence the computational outcome for grammatical analysis, because the number of 0-syntagma (utterances without a syntagma) is by default substracted from the number of utterances.

For interactional analysis on the other hand, which uses the turn and not the utterance as the basic unit, it is not of interest whether an interactional element is attached to one utterance or the other, unless it is within the same speaker turn.The decision whether to include an interactional element or any non-syntactical element into an utterance can follow the intonational pattern, and does not influence the grammatical analysis (which focuses on the syntagma alone) anymore.

An additional benefit is the re-definition of the utterance terminators.Because one utterance only contains one syntagma and vice versa, the utterance terminators .?! practically function as syntagma terminators. As such they reflect the sentence type of the syntagma: in other words, they define the syntagma as indicative, question, command, or exclamation. This can be done independently from the intonation, which can be represented by additional intonation markers -. -? -! preceding the terminator.

who did that -? ?
who did that -. ?

With separate intonation markers the intonation of questions with dislocated arguments (marked by„) can be represented as well, while also marking their status as question.

tabechatta no -?„ kore -.?

In the following we will try to apply this system, as well as to explore some of its analytical possibilities. The text (presented in full length in Appendix A) is based on a transcript cited by Terada (1994).

8. Transcription example

The transcription (Appendix A) uses the symbols explained above, as well as the usual CHAT symbols (MacWhinney 1995 [online version 1996/8]). By the use of the @ia, @is, @isc, and @if symbol for non-syntactics, it is possible to extract the syntactical structures produced (These structures which will be the target of any grammatical analysis,while the non-syntactic elements can be ignored [Appendix B]).
For example when applying the MLU program it is possible to exclude all words ending in@ia, @is, @isc, and @if, by using the option -s"*@i*" as in the following command (the +b+ option includes the + symbol, which is used as a compound symbol here, as a morpheme marker).

> mlu +b+ -s"*@i*" @

On the other hand it is possible to include the non-syntactical items when dispensing with the-s option.

> mlu +b+ @

The outcome (Appendix C) is astonishing at first glance. An MLU value of 7.750 for the L2 learner SIG is obtained when the non-syntactical items are included: a much

higher value of 10.800 is obtained, when they are excluded. This is explained by the fact, that the number of utterances is reduced, when the non-syntactical items are suppressed (56 vs 30). In other words, the high number of 0-syntagma (26 out of 56 utterances, as the substraction shows) causes a low MLU value. As 0-syntagma often consist of short strings, the inclusion of non-syntactical items lowers the MLU value. The morpheme number for the 26 0-syntagmas is 110, which would correspond to an MLU of 4.230, a still rather high value due to the frequent use of "soo＋desu＋ne". The same effect can be seen in the MLU value of the native speaker TOZ (4.167 vs 7.125).
On the other hand it is possible to focus on the interactional elements by using FREQ.

> freq ＋s%@i* ＋t*SIG ＋f @
> freq ＋s%@i* ＋t*TOZ ＋f @

By this simple frequency count (Appendix D) it becomes clear that the L2 learner SIG uses many more non-syntactic elements than the native speaker TOZ (SIG: 79, TOZ: 54), although the number of syntagmas is fairly equal (30 for SIG and 32 for TOZ, as we have seen in the MLU output). Moreover SIG uses a much higher number of speaker-inserted items (65 @is or 82%, 14 @ia or 18%), while the TOZ prefers auditor-inserted items (16 @is or 30%, 38 @ia or 70%).

The calculation of the ratio @i*/syntagma might even be a simple index for fluency. Here we get a ratio of 2.633 @i*/synt for SIG (number of syntagmas: 30, number of @i*: 79) and 1.688 @i*/synt for TOZ (number of syntagmas: 32, number of @i*: 54). The following commands yield a frequency list of the items used (Appendix E).

> freq ＋s*@i* ＋t*SIG ＋r4 ＋f @
> freq ＋s*@i* ＋t*TOZ ＋r4 ＋f @

For SIG a high number of non-syntactic elements (tokens) stands in contrast to their low variety (types), showing a low ttr (0.241). TOZ on the other hand displays a rather high ttr of 0.509. Actually except for the feedback signal "un", which occurs 17 times, most of his non-syntactical elements are produced only once or twice.

The vocabulary list can be obtained by the following commands (Appendix F).

> freq -s*@i* ＋t*SIG ＋r4 ＋r5 ＋f @
> freq -s*@i* ＋t*TOZ ＋r4 ＋r5 ＋f @

Also here the exclusion of the non-syntactic elements might help to get a clearer pic-

ture of the actual vocabulary of the L2 learner (131 types, ttr 0.510 for SIG, compared to 119 types, with a ttr of 0.688 for TOZ).

Notes

(1) Here lies an important difference to the T-unit (Hunt 1965) as well as C-unit (Loban 1977) concept, which only acknowledge clauses with co-referential deletion as belonging to the same unit, and not recognizing a connection only by "but", "and", or "or".

(2) For many items the orthographic status is weak, if they are graphically presented in written language at all, and the pronounciation is floating)

(3) This definition follows Trask (1993) who defines paralanguage as "the use of nonverbal elements in speech, such as intonation, expression and gestures in such a way as to affect the meaning of an utterance."

(4) In contrast to 'auditor backchannel responses' (Duncan / Fiske 1985: 58f.) which include syntagmatic elements like sentence completion, clarification requests and brief restatements as well)

(5) With reference to the terminus sociocentric sequence proposed by Bernstein (1962)

(6) "Ausdruecke, die, ausgezeichnet durch ihre Haeufigkeit in standardisierten Kommunikationssituationen, spezifische Formen fuer die Interaktion der Gespraechspartner uebernehmen".

(7) This can be changed optically by applying the SLIDE program. However this doesn't change the structure of the transcript

Cited Literature

Bernstein, Basil B. 1962.
Social class, linguistic codes and grammatical elements.
In: Language and Speech 5, 211–240

Blake, Joanna / Quartaro, Georgia / Onorati, Susan. 1993.
Evaluating quantitative measures of grammatical complexity in spontaneous speech samples.
In: Journal of Child Language 20, 139–152

Bloom, Lois. 1973.
One word at a time.
The Hague: Mouton

Clancy, Patricia M. 1982.
Written and spoken style in Japanese narratives.
In: Tannen, Deborah (ed). Spoken and written language. Exploring orality and literacy (vol.9).
Norwood: Abley P.C., 55–76

Coulmas, Florian / Marui, Ichiro / Reinelt, Rudolf. 1983.
Kleines Formellexikon Japanisch-Deutsch.
Berlin: E.Schmidt V.

Du Bois, John W. / Schuetze-Coburn, Stephan. 1993.

Representing hierarchy: constituent structure for discourse databases.
In: Edwards, Jane A. / Lampert, Martin D. (eds). Talking Data: transcription and coding in discourse research.
Hillsdale: LEA, 221–260

Duncan, Starkey / Fiske, Donald W. 1985.
The turn system.
In: Duncan, Starkey / Fiske, Donald W. eds. Interaction structure and strategy.
Cambridge: Cambridge U.P. 43–64

Edwards, Jane A. 1993.
Principles and contrasting systems of discourse transcription.
In: Edwards, Jane A. / Lampert, Martin D. (eds). Talking Data: transcription and coding in discourse research.
Hillsdale: LEA, 3–32

Huang, James C.-T. 1984
On the distribution and reference of empty pronouns
Linguistic Inquiry 15, 4, 531–574

Hunt,. 1970.
Syntactic maturity.
Society for research in Child Development Monographs 134

MacWhinney, Brian. 1995.
The CHILDES project: Tools for analyzing talk. 2nd ed.
Hillsdale, NJ: LEA

Oshima-Takane, Yuriko / MacWhinney, Brian (eds). 1995.
Nihongo no tame no CHILDES manyuaru.
Montreal: McGill University

Scott, Cheryl M. 1988.
Spoken and written syntax.
In: Nippold, Marilyn A. ed. Later language development: ages nine through nineteen.
Austin: proed. 49–95

Terada, Hiroko. 1994.
Nihongo no dainigengo shutoku judankenkyu to deetabeesuka ni tsuite no ikkosatsu [Observations concerning the acquisition of Japanese as second language and the transformation to a database].
In: Nihongo kenshuukoosu shuryosei tsuikachosa hokokusho.
Nagoya Daigaku Ryugakusei Sentaa. 160–187

Trask, Robert.L. 1993.
A dictionnary of grammtical terms in linguistics.

London: Routledge

Acknowledgements

I would like to express my gratitude to the members of theNihongo Kenshuukoosu Shuuryoosee Tsuiseki Choosa Project under Akito Ozaki (Nagoya University, Education Center for International Students), who allowed me insight to their research and the problems encoutered with data transcription, and generously gave me access to their valuable speech data. My special thanks go to Hiromi Morikawa (University of Kansas, Child Language Program), Craig Paul (University of Kansas), and Beverley Curran (Aichi Shukutoku J.C.) Their help and encouragement have largely contributed to this article.

<table>
<tr><td rowspan="9">interaction flow</td><td rowspan="8">utterances</td><td rowspan="2">syntagma</td><td colspan="3">sentences</td></tr>
<tr><td colspan="3">syntactically connected strings [ill-formed, uncompleted,elliptic clauses]</td></tr>
<tr><td rowspan="6">non-syntactic elements</td><td rowspan="5">interactional elements</td><td rowspan="2">auditor-inserted</td><td>auditor backchannel signals</td></tr>
<tr><td>paralanguage</td></tr>
<tr><td rowspan="3">speaker-inserted</td><td>speaker continuation signals</td></tr>
<tr><td>calling expressions</td></tr>
<tr><td>elliptic responses</td></tr>
<tr><td>paralanguage</td></tr>
<tr><td colspan="3">sociocentric formulas [high-frequency standardized social expressions]</td></tr>
<tr><td colspan="4">protosyntagma [one-word-utterances]</td></tr>
<tr><td colspan="5">non-interactional elements [coughing, hiccup, accent, body movements, etc]</td></tr>
</table>

Table 1. Scheme for the transcription of the interaction flow

Appendix A Transcribed and morphemicized text example

```
@Begin
@Participants:   SIG Singh Student, TOZ Ozaki Teacher
@Sex of SIG:     male
@Sex of TOZ:     male
@Age of SIG:     32;
@Language of SIG:     India [first language: ?]
@Date:           13-DEC-1993
@Situation:      free conversation
@Filename:       Sing1.92.9
@Coding:         JCHAT 1.0 Hebon 96/8
@Comment:        in order to be able to compare MLU values for both speakers,
          replacement brackets containing morphemicized words are used;
          non-syntactic elements are marked by special work markers
          atmark plus i (and derivations), and are excluded from MLU counting.
*TOZ:     sugoku [: sugoi-ku ] isogashisoo [: isogashii-soo ] ne: .
*SIG:     soo+desu+ne:@is &hehehe [=! warai] .
*TOZ:     yaa@is hora@is tegami morattara [: morau-tara ] ano@is getsuyoobi no go+ji to # nanka@is
          ni+kai gurai shika ne +...
*SIG:     soo+desu+ne@ia -? .
*TOZ:     +, moo hima-na jikan ga nai tte kaiteatta [: kaku-te+aru-ta ] deshoo.
*SIG:     soo+desu+ne@ia .
*TOZ:     are mite [: miru-te ] bikkuri shichatte [: suru-chau-te] +...
*SIG:     ni(+kai) [/] ni+kai toka san+kai kanaa # to +...
*TOZ:     aa+soo+ka@ia -. ?
*TOZ:     yappa(ri) isogashii ?
*SIG:     soo+desu+ne:@is .
*SIG:     aa@is itsumo anoo@is # jikken anoo@is asa kara anoo@is yoru tabun juu+ichi+ji made ni mo ikimasu
          [: iku-masu ] [*] .
%err:     ikimasu = narimasu $LEX
*TOZ:     aa@ia .
*SIG:     tokidoki anoo@is # owannai [: owaru-nai ] toki xx desu .
*TOZ:     a+soo@ia -. ?
*SIG:     hai@is .
*TOZ:     to o#yasumi nanka wa doo sun [: suru ] no -? +/?
*SIG:     o#yasumi < ni wa > [>] +/.
*TOZ:     +, ,, < fuyu+yasumi > [<] toka natsu+yasumi toka saa -. ?
*SIG:     anoo@is # gakusee-tachi wa yasumi desu keredomo ano@is boku-tachi wa ano@is +/.
*TOZ:     a+soo+ka@ia -. ?
*SIG:     +, yasumi +...
*TOZ:     moo [/] moo gakusee tte yuu yori wa kenkyuusha nano ne ?
*SIG:     kenkyuusho +...
*TOZ:     soo@ia .
*SIG:     +, [//] kenkyuusee .
%com:     selfcorrection
*SIG:     un@is ano@is # nichiyoobi ni mo ano@is +/.
*TOZ:     +^ kiteru [: kuru-teru ] .
*SIG:     +, kiteru [//] kiteimasu [: kuru-te+iru-masu ] +/.
*TOZ:     hoo@ia .
*SIG:     un@is .
*SIG:     +, ,, ano@is # hayai [*] nanika yaritai [: yaru-tai ] desu kara .
%err:     hayai = hayaku $MOR
*TOZ:     un@ia .
*TOZ:     kono aida wa denwa shita [: suru-ta ] toki saa +/.
*SIG:     hai@ia .
*TOZ:     +, Shingu-san ga deta [: deru-ta ] deshoo ?
*SIG:     soo+desu+ne@ia -? [*] .
%err:     soo+desu+ne@ia  = soo+desu+ne@ia $PHO
%com:     rising intonation instead of falling
*TOZ:     boku ne: ryuugakusee da to omowanakatta [: omou-nai-ta ] .
*SIG:     a+soo+desu+ka@ia -. ?
*SIG:     &hahaha [=! warai ] .
*TOZ:     un@is .
*TOZ:     &hahaha [=! warai ] .
*TOZ:     iya@is un@is sorede ano@is < Shingu-san ni hanashitai [: hanasu-tai ] ndesu kedo > ["] tte
          ittara [: iu-tara ] +/.
*SIG:     hai@ia .
*TOZ:     +, ano@is < Shingu desu > ["] tte iwarete [: iu-rareru-te ] eeQ ["] tte omotta [>]
          [: omou-ta]  .
*TOZ:     < are wa bikkuri > [>] shita [: suru-ta ] ,, hontoo ni .
*SIG:     aa+soo+desu+ka@ia [<] -. ?
*SIG:     aa+soo+desu+ka@ia -. ?
*TOZ:     un@is .
*SIG:     un@is .
*TOZ:     jo(ozu) [/] joozu n(i) natchatta [: naru-chau-ta ] njanai ,, nihongo -. ?
*SIG:     umm@is .
*SIG:     demo ano@is moo [*] ano@is benkyoo shitai [: suru-tai ] desu .
%err:     moo = motto [?] $LEX
```

```
*SIG:    demo chotto jikan ga # nai desu kara komarimasu [: komaru-masu ] ne: .
*TOZ:    ano@is rokkagetsu [: roku+kagetsu ] owatte [: owaru-te ] sa(a) +...
*SIG:    hai@ia .
*TOZ:    +, daitai ichi+nen chotto &deto [?] desho ?
*SIG:    soo@is ichi+nen+han kurai kana(a) ?
*TOZ:    ichi+nen+han kurai kana(a) ?
*SIG:    soo+desu+ne@ia -? [*] .
%err:    soo+desu+ne@ia  = soo+desu+ne@ia $PHO
%com:    rising intonation instead of falling
*TOZ:    de itsu+goro kara jibun de < aa@is nihongo mo [?] moo daijoobu da > tte omou no ka ne: -. ?
*SIG:    itsu kara -. ?
*TOZ:    un@is .
*TOZ:    datte rokkagetsu [: roku+kagetsu ] owatta [: owaru-ta ] toki sa(a) +/.
*SIG:    soo+desu+ne@is .
*SIG:    ano@is # sono toki wa ano@is # hanashi ga # aa@is # denakatta [: deru-nai-ta ] desu ne -? .
*TOZ:    un+un+un@ia .
*SIG:    demo # ano@is # kenkyuu+shitsu ni iku [*] toki wa ne ano@is # mina-san ga ano@is
         nihonjin+gakusee toka +/.
*TOZ:    un@ia .
*SIG:    +, sensee-gata mo +/.
*TOZ:    un@ia .
*SIG:    ano@is itsumo nihongo de hanashimasu [: hanasu-masu ] .
*TOZ:    un+un@ia .
*SIG:    sensee-gata wa tokidoki # eego de gambarimasu [: gambaru-masu ] .
*TOZ:    &huhuhu [=! warai] .
*SIG:    &ehehe [=! warai] .
*TOZ:    eego de gambaru wake -. ?
*TOZ:    &hahaha [=! warai] .
*SIG:    hai@is .
*SIG:    &hahaha [=! warai] .
*SIG:    demo ano@is # honto(o) ni komarimasu [: komaru-masu] ne: .
*TOZ:    un@ia .
*SIG:    sono toki wa ano@is # boku-tachi wa ano@is gamabaranaito [: gambaru-nai-to ] ikenai omotte
         [: omou-te ] [*] ano@is itsumo nanika ano@is atarashii kotoba kiku toki wa <ja nan to yuu imi
         desu ka > ["] < doo yatte [: yaru-te ] puronaunsu [= hatsuon] shimasu [: suru-masu ]
         ka > ["] +/.
%err:    omotte = to_omotte $MOR
*TOZ:    un+un+un@ia .
*SIG:    +, toka &i:aite [?] < kanji no [*] nan to yomu ndesu ka > ["] toka iroiro # um@is # nihonjin no
         gakusee ni kiite [: kiku-te ] tetsudattemoratteimasu [: tetsudau-te+morau-te+iru-masu ] .
%err:    no = de  [?] $LEX
*TOZ:    un@ia .
*SIG:    soo+desu@is .
*TOZ:    fuun@ia .
*TOZ:    dakedo sono # rokkagetsu [: roku+kagetsu ] owatte [: owaru-te ] +/?
*SIG:    un@ia .
*TOZ:    +, ma(a)@is ikkagetsu [: ichi+kagetsu ] ni+kagetsu koo dandan tatte [: tatsu-te ] +/?
*SIG:    soo+desu+ne@ia .
*TOZ:    +, nan+kagetsu gurai tatsu-to ne Shingu-san no baai wa itsu+goro kara aa@is +/?
*SIG:    soo+desu+ne@is .
*TOZ:    un@ia .
*SIG:    tabun san+yon+kagetsu ato gurai kamoshirenai .
*TOZ:    aa+soo@ia .
*SIG:    soo+desu@is .
*TOZ:    fuun@ia .
*SIG:    de ano@is Sensee@isc ato wa ne -? +/.
*TOZ:    un@ia .
*SIG:    +, ano@is hitotsu happyoo ga arimashita [: aru-masu-ta ] .
*SIG:    ano@is # sono happyoo wa nihongo de happyoo +/.
*TOZ:    +^ suru .
*SIG:    ano@is Ryuugakusee+sentaa +/.
*TOZ:    un@ia .
*SIG:    +, Minato-ku +/.
*TOZ:    haa@ia .
*SIG:    +, ni [*] # ano@is tabun mainen [//] ano@is kyonen kara hajimatte [: hajimaru-te ] +/.
%err:    ni = de $LEX
*TOZ:    un@ia .
*SIG:    ima [*] ano@is sugu [*] mainen yaru to omoimasu [: omou-masu ] .
%err:    ima = moo [?] ; sugu = [?]
*TOZ:    u@ia nani o yatta [: yaru-ta ] no ?
*SIG:    ano@is soko ni [*] wa +/.
%err:    ni = de $LEX
*TOZ:    un@ia .
*SIG:    +, boku no hanashi wa +/.
*TOZ:    un@ia .
*SIG:    +, ano@ia < nichijoo+seekatsu ni okeru kokusai+kooryuu to kokusai+rikai > ["] +/.
*TOZ:    hee@ia .
*TOZ:    < donna hanashi shita [: suru-ta] no > [>] ?
*SIG:    +, < ni tsuite hanashimashita [: hanasu-ta ] > [<] .
```

```
*SIG:    sore wa ano@is # iroiro # ano@is Nihon de wa < nan to > [//] nan desu ka +/.
*TOZ:    un@ia .
*SIG:    +, to # gaikoku de wa +/.
*TOZ:    un@ia .
*SIG:    ano@ia # < doo yatte > [*] chigaimasu [: chigau-masu ] ka toka doo shitara [: suru-tara ] ii
         desu ka toka iroiro hanashimashita [: hanasu-masu-ta ] .
%err:    doo_yatte = dono_yoo_ni $MOR
*TOZ:    un@ia .
*TOZ:    nanpun [: nan+fun ] gurai ?
*SIG:    hachipun@n [//] happun [: hachi+fun] .
*TOZ:    supiichi mitai ?
*SIG:    supiichi .
*TOZ:    haan@ia .
*SIG:    ano@is # soko de wa anoo@is # taado+pureesu [= third place] [//] san'i [>] +...
*TOZ:    aa+soo+datta+no@ia [<] -. ?
*SIG:    hai@is .
*TOZ:    ja nanika moratta [: morau-ta ] ?
*SIG:    ano@is # ni+man+en # +/.
*TOZ:    aa+soo@ia -. ?
*SIG:    +, moraimashita [: morau-masu-ta ] .
*TOZ:    fuun@ia .
*TOZ:    a@is de sore yatte [: yaru-te ] yokatta [: ii-ta ] ?
*SIG:    soo+desu@is .
*SIG:    ano@is # ano@is # ni+hyaku+nin gurai nihonjin ga +/.
*TOZ:    un@ia .
*SIG:    +, ano@is # kimashita [: kuru-masu-ta ] .
*TOZ:    a+soo+desu+ka@ia -. ?
*SIG:    de ano@is # shimbun de mo dekimashita [: dekiru-masu-ta ] [*] .
%err:    dekimashita = demashita $LEX
*TOZ:    ara@ia .
*TOZ:    sore shirankatta [: shiru-nai-ta ] naa .
*SIG:    ano@is # < boku no baai > [*] wa +/.
%err:    boku_no_baai = boku_ni_tsuite $LEX
*TOZ:    un@ia .
*SIG:    +, ano@is sambunno'ichi [: san+bun+no+ichi ] gurai +/.
*TOZ:    un@ia .
*SIG:    +, ano@is kakimashita [: kaku-masu-ta ] [*] .
%err:    kakimashita = kaitearimashita $MOR
*TOZ:    shimbun ni ?
*SIG:    shimbun ni ano@is # zembu no sambunno'ichi [: san+bun+no+ichi ] gurai boku no hanashi ga
         ano@is ## +...
*TOZ:    +^ deteta [: deru-teru-ta ] .
*SIG:    soo+desu@is .
@End
```

Appendix B List of the Syntagma

```
@Begin
@Participants:    SIG Singh Student, TOZ Ozaki Teacher
@Coding:    JCHAT 1.0 Hebon
@Comment:    non-syntactic elements excluded.
        repetitions and selfcorrections excluded
        unmorphemicized (no morphemic replacement)
        no intonation, no overlap marked

*TOZ:    sugoku isogashisoo ne: .
*TOZ:    tegami morattara getsuyoobi no goji to  nikai gurai
        shika ne moo himana jikan ga nai tte kaiteatta deshoo .
*TOZ:    are mite bikkuri shichatte +...
*SIG:    nikai toka sankai kanaa to +...
*TOZ:    yappa(ri) isogashii ?
*SIG:    itsumo jikken asa kara yoru tabun juuichiji made ni mo ikimasu [*] .
*SIG:    tokidoki owannai toki xx desu .
*TOZ:    to oyasumi nanka wa doo sun no ,, fuyuyasumi toka
        natsuyasumi toka saa ?
*SIG:    oyasumi ni wa +...
%eng:    in the vacation +...
*SIG:    gakuseetachi wa yasumi desu keredomo bokutachi wa yasumi +...
*TOZ:    moo gakusee tte yuu yori wa kenkyuusha nano ne ?
*SIG:    kenkyuusee .
*SIG:    nichiyoobi ni mo kiteimasu ,, hayai [*] nanika yaritai desu kara .
*TOZ:    +^ kiteru [: nichiyoobi ni mo kiteru ] .
*TOZ:    kono aida wa denwa shita toki saa Shingusan ga deta deshoo ?
*TOZ:    boku ne: ryuugakusee da to omowanakatta .
*TOZ:    sorede < Shingusan ni hanashitai ndesu kedo > ["] tte ittara < Shingu
        desu > ["] tte iwarete eeQ ["] tte omotta .
*TOZ:    are wa bikkuri shita ,, hontoo ni .
*TOZ:    joozu n(i) natchatta njanai ,, nihongo ?
*SIG:    demo moo [*] benkyoo shitai desu .
*SIG:    demo chotto jikan ga nai desu kara komarimasu ne: .
*TOZ:    rokkagetsu owatte sa(a) daitai ichinen chotto desho ?
*SIG:    ichinenhan kurai kana(a) ?
*TOZ:    ichinenhan kurai kana(a) ?
*TOZ:    de itsugoro kara jibun de < nihongo mo [?] moo daijoobu da >
        tte omou no ka ne: ?
*SIG:    itsu kara ?
*TOZ:    datte rokkagetsu owatta toki sa(a) +...
*SIG:    sono toki wa hanashi ga denakatta desu ne .
*SIG:    demo kenkyuushitsu ni iku [*] toki wa ne minasan ga nihonjin+gakusee
        toka senseegata mo itsumo nihongo de hanashimasu .
*SIG:    senseegata wa tokidoki eego de gambarimasu .
*TOZ:    eego de gambaru wake ?
*SIG:    demo honto(o) ni komarimasu ne: .
*SIG:    sono toki wa bokutachi wa gamabaranaito ikenai omotte [*] itsumo
        nanika atarashii kotoba kiku toki wa <ja nan to yuu imi desu ka >
        ["] < doo yatte puronaunsu shimasu ka > ["] toka < kanji no [*] nan
        to yomu ndesu ka > ["] toka iroiro nihonjin no gakusee ni kiite
        tetsudattemoratteimasu .
*TOZ:    dakedo sono rokkagetsu owatte ikkagetsu nikagetsu koo dandan tatte
        nankagetsu gurai tatsuto ne Shingusan no baai wa itsugoro kara +...
*SIG:    tabun sanyonkagetsu ato gurai kamoshirenai .
*SIG:    de ato wa ne hitotsu happyoo ga arimashita .
*SIG:    sono happyoo wa nihongo de happyoo +...
*TOZ:    +^ suru [: sono happyoo wa nihongo de happyoo suru ] .
```

*SIG: Ryuugakusee+sentaa Minato-ku ni [*] tabun kyonen kara hajimatte ima [*] sugu [*] mainen yaru to omoimasu .
*TOZ: nani o yatta no ?
*SIG: soko ni [*] wa boku no hanashi wa < nichijoo+seekatsu ni okeru kokusai+kooryuu to kokusai+rikai > ["] ni tsuite hanashimashita .
*TOZ: donna hanashi shita no ?
*SIG: sore wa iroiro Nihon de wa nan desu ka to gaikoku de wa < doo yatte > [*] chigaimasu ka toka doo shitara ii desu ka toka iroiro hanashimashita .
*TOZ: nanpun gurai ?
*SIG: happun.
*TOZ: supiichi mitai ?
*SIG: supiichi .
*SIG: soko de wa san'i +...
*TOZ: ja nanika moratta ?
*SIG: niman'en moraimashita .
*TOZ: de sore yatte yokatta ?
*SIG: nihyakunin gurai nihonjin ga kimashita .
*SIG: de shimbun de mo dekimashita [*] .
*TOZ: sore shirankatta naa .
*SIG: < boku no baai > [*] wa sambunno'ichi gurai kakimashita [*] .
*TOZ: shimbun ni ?
*SIG: shimbun ni zembu no sambunno'ichi gurai boku no hanashi ga +...
*TOZ: +^ deteta [: shimbun ni zembu no sambunno'ichi gurai boku no hanashi ga deteta] .
@End

Appendix C MLU output

NON-SYNTACTICS EXCLUDED:

```
> mlu +b+ -s"*@i*" @
****************************************
```

+b+ for recognition of + as morpheme delimiter

-s"*@i*" for exclusion of all words * ending in @i plus something *

```
****************************************
From file <SIG.cha FORuttDef>
MLU for Speaker: *SIG:
  MLU (xxx and yyy are EXCLUDED from the utterance and morpheme counts):
    Number of: utterances = 30, morphemes = 324
        Ratio of morphemes over utterances = 10.800
        Standard deviation = 2.587

MLU for Speaker: *TOZ:
  MLU (xxx and yyy are EXCLUDED from the utterance and morpheme counts):
    Number of: utterances = 32, morphemes = 228
        Ratio of morphemes over utterances = 7.125
        Standard deviation = 3.029
```

NON-SYNTACTICS INCLUDED:

```
> mlu  +b+  @

****************************************

MLU for Speaker: *SIG:
  MLU (xxx and yyy are EXCLUDED from the utterance and morpheme counts):
    Number of: utterances = 56, morphemes = 434
        Ratio of morphemes over utterances = 7.750
        Standard deviation = 5.412

MLU for Speaker: *TOZ:
  MLU (xxx and yyy are EXCLUDED from the utterance and morpheme counts):
    Number of: utterances = 72, morphemes = 300
        Ratio of morphemes over utterances = 4.167
        Standard deviation = 3.930
```

Appendix D FREQ number of non-syntactical elements

```
freq +s%@i* +t*SIG +f @
  ONLY speaker main tiers matching: *SIG;
****************************************
 14 @ia
 64 @is
  1 @isc
------------------------------
    3  Total number of different word types used
   79  Total number of words (tokens)
0.038  Type/Token ratio


 number of utterances: 56
 number of syntagmas: 30
 number of @i*: 79
  65 @is =  82%
  14 @ia =  18%
  1.411 @i*/utt ratio for SIG
  2.633 @i*/synt ratio for SIG


freq +s%@i* +t*TOZ +f @
   ONLY speaker main tiers matching: *TOZ;
****************************************
 38 @ia
 16 @is
------------------------------
    2  Total number of different word types used
   54  Total number of words (tokens)
0.037  Type/Token ratio



 number of utterances: 72
 number of syntagmas: 32
 number of @i*: 54
  16 @is = 30%
  38 @ia = 70%
  0.750 @i*/utt ratio for TOZ
  1.688 @i*/synt ratio for TOZ
```

Appendix E FREQ list of non-syntactical elements

```
> freq +s*@i* +t*SIG +r4 +f @
                          +r4 to remove : elongation symbol from words
                          +t*SIG  to analyse only utterances of SIG
  ONLY speaker main tiers matching: *SIG;
****************************************

  1 a+soo+desu+ka@ia
  2 aa+soo+desu+ka@ia
  2 aa@is
  2 ano@ia
 39 ano@is
  6 anoo@is
  3 hai@ia
  3 hai@is
  1 sensee@isc
  5 soo+desu+ne@ia
  4 soo+desu+ne@is
  4 soo+desu@is
  1 soo@is
  1 um@is
  1 umm@is
  1 un@ia
  3 un@is
------------------------------
   17  Total number of different word types used
   79  Total number of words (tokens)
0.215  Type/Token ratio

> freq +s*@i* +t*TOZ +r4 +f @
 ONLY speaker main tiers matching: *TOZ;
****************************************
  1 a+soo+desu+ka@ia
  1 a+soo+ka@ia
  1 a+soo@ia
  1 a@is
  1 aa+soo+datta+no@ia
  1 aa+soo+ka@ia
  2 aa+soo@ia
  1 aa@ia
  2 aa@is
  4 ano@is
  1 ara@ia
  3 fuun@ia
  1 haa@ia
  1 haan@ia
  1 hee@ia
  1 hoo@ia
  1 hora@is
  1 iya@is
  1 maa@is
  1 nanka@is
  1 soo@ia
  1 u@ia
  2 un+un+un@ia
  1 un+un@ia
 17 un@ia
  4 un@is
  1 yaa@is
------------------------------
   27  Total number of different word types used
   54  Total number of words (tokens)
0.500  Type/Token ratio
```

Appendix F FREQ vocabulary list (without non-syntactical elements)

```
> freq -s*@i* +t*SIG +r4 +r5 +f @
                              +r4  to remove elongation symbol : from words
                              +r5  to stop replacement  [: ]
  ONLY speaker main tiers matching: *SIG;
****************************************
  1 arimashita
  1 asa
  1 atarashii
  2 ato
  1 baai
  1 benkyoo
  3 boku
  2 boku-tachi
  1 chigaimasu
  1 chotto
  9 de
  1 dekimashita
  4 demo
  1 denakatta
  9 desu
  3 doo
  1 eego
  6 ga
  1 gaikoku
  1 gakusee
  1 gakusee-tachi
  1 gamabaranaito
  1 gambarimasu
  4 gurai
  1 hachipun@n
  1 hajimatte
  3 hanashi
  2 hanashimashita
  1 hanashimasu
  1 happun
  3 happyoo
  1 hayai
  1 hitotsu
  1 hontoo
  1 ichi+nen+han
  1 ii
  1 ikenai
  1 ikimasu
  1 iku
  1 ima
  1 imi
  3 iroiro
  1 itsu
  3 itsumo
  1 ja
  1 jikan
  1 jikken
  1 juu+ichi+ji
  6 ka
  1 kakimashita
  1 kamoshirenai
  2 kanaa
  1 kanji
  5 kara
  1 kenkyuu+shitsu
  1 kenkyuusee
  1 kenkyuusho
  1 keredomo
  1 kiite
  1 kiku
  1 kimashita
  1 kiteimasu
  1 kiteru
  1 kokusai+kooryuu
  1 kokusai+rikai
  2 komarimasu
  1 kotoba
  1 kurai
```

```
  1 kyonen
  1 made
  2 mainen
  1 mina-san
  1 minato-ku
  4 mo
  1 moo
  1 moraimashita
  1 nai
  4 nan
  2 nanika
  1 ndesu
  5 ne
 11 ni
  1 ni+hyaku+nin
  2 ni+kai
  1 ni+man+en
  1 nichijoo+seekatsu
  1 nichiyoobi
  1 nihon
  2 nihongo
  2 nihonjin
  1 nihonjin+gakusee
  6 no
  1 o#yasumi
  1 okeru
  1 omoimasu
  1 omotte
  1 owannai
  1 puronaunsu
  1 ryuugakusee+sentaa
  2 sambunno'ichi
  1 san'i
  1 san+kai
  1 san+yon+kagetsu
  2 sensee-gata
  1 shimasu
  2 shimbun
  1 shitai
  1 shitara
  2 soko
  3 sono
  1 sore
  1 sugu
  1 supiichi
  1 taado+pureesu
  3 tabun
  1 tetsudattemoratteimasu
  7 to
  6 toka
  5 toki
  2 tokidoki
  1 tsuite
 18 wa
  1 xx
  1 yaritai
  1 yaru
  2 yasumi
  2 yatte
  1 yomu
  1 yoru
  1 yuu
  1 zembu
------------------------------
  131  Total number of different word types used
  257  Total number of words (tokens)
0.510  Type/Token ratio


> freq -s*@i* +t*TOZ +r4 +r5 +f @
  ONLY speaker main tiers matching: *TOZ;
****************************************
  1 aida
  2 are
  1 baai
  2 bikkuri
  1 boku
```

1 chotto
2 da
1 daijoobu
1 daitai
1 dakedo
1 dandan
1 datte
4 de
1 denwa
1 desho
2 deshoo
1 desu
1 deta
1 deteta
1 donna
1 doo
1 eego
1 eeq
1 fuyu+yasumi
2 ga
1 gakusee
1 gambaru
1 getsuyoobi
1 go+ji
3 gurai
1 hanashi
1 hanashitai
1 hima-na
1 hontoo
1 ichi+nen
1 ichi+nen+han
1 ikkagetsu
1 isogashii
1 isogashisoo
2 itsu+goro
1 ittara
1 iwarete
1 ja
1 jibun
1 jikan
2 joozu
1 ka
1 kaiteatta
1 kanaa
2 kara
1 kedo
1 kenkyuusha
1 kiteru
1 kono
1 koo
1 kurai
1 mitai
1 mite
1 mo
4 moo
1 moratta
1 morattara
1 naa
1 nai
1 nan+kagetsu
1 nani
1 nanika
1 nanka
1 nano
1 nanpun
1 natchatta
1 natsu+yasumi
1 ndesu
6 ne
4 ni
1 ni+kagetsu
1 ni+kai
2 nihongo
1 njanai
6 no
1 o
1 o#yasumi
1 omotta
1 omou

```
  1 omowanakatta
  1 owatta
  2 owatte
  3 rokkagetsu
  1 ryuugakusee
  4 saa
  1 shichatte
  1 shika
  1 shimbun
  1 shingu
  3 shingu-san
  1 shirankatta
  3 shita
  1 sono
  2 sore
  1 sorede
  1 sugoku
  1 sun
  1 supiichi
  1 suru
  1 tatsu-to
  1 tatte
  1 tegami
  3 to
  2 toka
  2 toki
  6 tte
  5 wa
  1 wake
  1 yappari
  1 yatta
  1 yatte
  1 yokatta
  1 yori
  1 yuu
------------------------------
  119  Total number of different word types used
  173  Total number of words (tokens)
0.688  Type/Token ratio
```