Syntactic–prosodic boundary labels for English:

Annotation, correlation, classification.

(preliminary version 0.5, comments welcome)

Anton Batliner, Jan Buckow,
Richard Huber, Volker Warnke

Abstract and disclaimer

This paper is a sequel of [**?**]. We describe the M labels for English that are slightly modified with respect to the German version, their correlation with acoustic–perceptual boundary labels and with dialogue act boundary labels as well as recognition results. This version is preliminary, a knowledge of [**?**, **?**, **?**] will make understanding much easier. Some passages are for the moment rather sketchy.

label	description
type	**S**entence, free **P**hrase **L**eft dislocation, **R**ight dislocation **E**mbedded sentence/phrase **D**iscourse particle (**D**, **T**) **A**mbiguous boundary **I**nternal constituent boundary
hierarchy	**M**ain, **S**ubordinate, **C**oordinate
strength	prosodic–syntactic strength: strong (3), intermediate (2), weak (1), very weak (0)

Table 1: Description of labels

1 Introduction

The revised version of the M labelling scheme for German is described in [**?**] and, within a larger context, in [**?**]. In the meantime, the native English dialogues of VM–CD 6 and all dialogues of VM–CD's 8 and 13 are annotated with a slightly modified version of the German M labels. For a thorough understanding, it might be necessary to read one of these papers 'synoptically' together with the present one; [**?**] is shorter but a more sketchy, [**?**] gives an outlook of the whole endeavour but is, because of that, much longer. [**?**] is detailed as well but does only deal with the first version of the M labelling scheme.

2 The Labelling System

The names of the labels consist of three characters each with the following encoding:

For type, we use fairly well–known terms. Note, however, that the extensional and intensional definition can change across linguistic theories. With strength, we encode a mixture, mainly of prosodic, but partly of syntactic, strength. This is at the same time our working hypothesis that prosodic and syntactic strength covary to a great extent,

For the convenience of the readers, the new German M labels are listed in Tables 1 and 2, where the mapping onto the old labels, the context with one example for each label, the label itself, and the main class it is attached to are given; these tables can be found in [**?**] as well. The names of the new labels consist of three characters each with the encoding given in Table 3. Type and hierarchy describe syntactic phenomena; with strength, we so to speak code our working hypothesis that prosodic (and thereby, to a lesser extent, syntactic) marking of boundaries is scaled along these lines. Most of the revisions concern a sub–specification of the former M labels that most of the time could not take into account

hierarchical dependencies and left/right relationship.

2.1 A feature matrix for the M labels

In the feature matrix of Table **??**, the following features are displayed:

feature	description
sentence	functioning as a sentence ('satzwertig')
verb	with verb ('satzförmig')
left	attributed/subordinated to the left
right	attributed/subordinated to the right
ambiguous	(syntactically/semantically) ambiguous

In the following, the meaning of these features is described shortly:

sentence: The sequence of words (chunk) in question is functioning the same way (or is) a 'normal' sentence including elliptic sentences (free phrases); it is 'satzwertig', i.e., it shows the behavior and function of a 'normal' sentence.

verb: The sequence of words (chunk) in question contains a verb (finitum or infinitum); i.e. normally, it has not only the function of a sentence but its form as well ('satzförmig'); clauses e.g. are both [+sentence] and [+verb].

left/right related: This relationship is either hierarchical ore purely linear: Subordinate clauses are attached to their matrix sentence, coordinated, partly elliptic main clauses are attached to the adjacent non-elliptic main clause. Dislocated phrases that typically are referred to in the adjacent clause with a pro element are attached to this clause. On the other hand, non coordinated main clauses and free phrases are not related to the left or right. Particles that we take into account are either presentential and thus attached to the clause to their right or postsentential and thus attached to the clause to their left. Note that this feature is not always unequivocal.

ambiguous: These boundaries represent possible syntactic boundary positions. In our context, the alternative interpretation that they trigger are however normally not only purely syntactic but semantic or functional as well.

2.2 A short characterization of the label classes

Generally, we do not want to sub-specify beyond the levels given by our features, i.e., we cannot specify two levels of subordination. Other possible sub-specification are merged, e.g., if an elliptic sentence (free phrase) is followed by a subordinated sentence, we label this boundary with SM2; this constellation is very rare, and because of that, it makes not much sense to model it especially.

main class	label	context (between/at) with example
sentences, up to now: M3S		
M3	SM3	**Main clause and main clause:** *vielleicht stelle ich mich kurz vorher noch vor SM3 mein Name ist Lerch* perhaps I should first introduce myself SM3 my name is Lerch
M3	SM2	**Main clause and subordinate clause:** *ich weiß nicht SM2 ob es auch bei Ihnen dann paßt* I don't know SM2 whether it will suit you or not
M3	SS2	**Subordinate clause and main clause:** *da ich aus Kiel komme SS2 wird hier ja relativ wenig gefeiert* because I am from Kiel SS2 we don't celebrate that often
M3	SM1	**Main clause and subordinate clause, prosodically integrated:** *ich denke SM1 das können wir so machen* I think SM1 we can do it that way
M3	SS1	**Subordinate clause and main clause, prosodically integrated:** *das sieht sowieso ziemlich schlecht aus SS1 würd' ich sagen* anyway, that looks rather bad SS1 I'd say
M3	SC3	**Coordination of main clauses and of subordinate clauses:** *dann nehmen wir den Montag SC3 und treffen uns dann morgens* then we'll take Monday SC3 and meet in the morning
M3	SC2	**Subordinate clause and subordinate clause:** *da ich froh wäre SC2 diese Sache möglichst schnell hinter mich zu bringen* because I would be glad SC2 to get it over as soon as possible
free Phrases, up to now: M3P		
M3	PM3	**free Phrase, stand alone:** *sehr gerne PM3 ich liebe Ihre Stadt* with pleasure PM3 I love your town
M2	PC2	**sequence in free Phrases:** *um neun Uhr PC2 in 'nem Hotel PC2 in Stockholm* at nine o'clock PC2 in a hotel PC2 in Stockholm
M3	PM1	**free Phrase, prosodically integrated, no dialogue act boundary:** *guten Tag PM1 Herr Meier* hello PM1 Mr. Meier
Left dislocations, up to now: M3P		
M3	LS2	**Left dislocation:** *am fünften LS2 da hab' ich etwas* on the fifth LS2 I am busy
M2	LC2	**sequence of Left dislocations:** *aber zum Mittagessen LC2 am neunzehnten LS2 wenn Sie vielleicht da Zeit hätten* but for lunch LC2 on the 19th LS2 if you've got time then
Right dislocations, up to now: M3E		
M3	RS2	**Right dislocation:** *wie würde es Ihnen denn am Dienstag passen RS2 den achten Juni* will Tuesday suit you RS2 the eighth of June
M2	RC2	**sequence of Right dislocations:** *es wäre bei mir dann möglich RS2 ab Freitag RC2 dem fünfundzwanzigsten* it would be possible for me RS2 from Friday onwards RC2 the 25th
M2	RC1	**Right 'dislocation' at open verbal brace:** *treffen wir uns RC1 um eins* let's meet RC1 at one o'clock

Table 2: Examples for new boundary labels and their context, part I.

main class	**label**	**context (between/at) with example**
		Embedded strings, up to now: M3I
M3	EM3	**Embedded sentence/phrase:** *eventuell EM3 wenn Sie noch mehr Zeit haben EM3 <Atmung> 'n bißchen länger* possibly EM3 if you've got even more time <breathing> EM3 a bit longer
		Free particles, up to now: M3T
M3	FM3	**pre-/postsentential particle, with <pause> etc.:** *gut FM3 <Pause> okay* fine FM3 <pause> okay
		Discourse particles, up to now: M3D
MU	DS3	**pre-/postsentential particle, ambisentential:** *dritter Februar DS3 ja DS3 ab vierzehn Uhr hätt' ich da Zeit* third February DS3 isn't it/well DS3 I have time then after two p.m.
MU	DS1	**pre-/postsentential particle, no <pause> etc.:** *also DS1 dienstags paßt es Ihnen DS1 ja M3S <Atmung>* then DS1 Tuesday will suit you DS1 won't it / after all <breathing>
		Ambiguous boundaries, up to now: M3A
MU	AM3	**between sentences, Ambiguous:** *würde ich vorschlagen AM3 vielleicht AM3 im Dezember AM3 noch mal AM3 dann* I'd propose AM3 possibly AM3 in December AM3 again AM3 then
MU	AM2	**between free phrases, Ambiguous:** *sicherlich AM2 sehr gerne* sure/-ely AM2 with pleasure
MU	AC1	**between constituents, Ambiguous:** *wollen wir dann AC1 noch AC1 'n Treffen machen* should we then (still) have a meeting / should we then have another meeting
		Constituents, up to now: M2I
M2	IC2	**between Constituents:** *ich wollte gerne mit Ihnen IC2 ein Frühstück vereinbaren* I'd like to arrange IC2 a breakfast with you
M2	IC1	**asyndetic listing of Constituents (not labelled up to now):** *wir haben bis jetzt eins IC1 zwei IC1 drei IC1 vier IC1 fünf IC1 sechs Termine* until now, we've got one IC1 two IC1 three IC1 four IC1 five IC1 six appointments
		default, no boundary, up to now: M0
M0	IC0	**every other word boundary:** *da bin ich ganz Ihrer Meinung* I fully agree with you

Table 3: Examples for new boundary labels and their context, part II.

label	description
type	**S**entence free **P**hrase **L**eft dislocation **R**ight dislocation **E**mbedded sentence/phrase **F**ree particle **D**iscourse particle **A**mbiguous boundary **I**nternal constituent boundary
hierarchy	**M**ain, **S**ubordinate, **C**oordinate
strength	prosodic–syntactic strength: strong (3), intermediate (2), weak (1), very weak (0)

Table 4: Encoding of type, hierarchy, and strength.

2.3 Sentences: S

For this class, we denote subordination, coordination, left/right relationship and prosodic marking. With these distinctions, we cannot denote all **all** constellations. E.g., we only have one level for subordination, i.e., with SC2, we cannot denote which one of these clauses is subordinated w.r.t the other one. After free phrases (elliptic sentences) followed by a subordinate clause, SM2 or SM1 is labelled as well: *" Wunderbar* SM1 *daß Sie da Zeit haben."* Analogously, phrasal coordination at subordinate clauses is labelled with SC3.

2.4 Phrases: P

Besides the 'main' label PM3, we annotate free phrases that are prosodically integrated with the following adjacent sequence with PM1. Sequences inside free phrases are analogous to the constituent boundaries IC2 and labelled with PC2.

2.5 Left dislocations: L

Left dislocations are constituents to the left of the matrix sentence, typically but not necessarily with some sort of anaphoric reference in the matrix sentence.

2.6 Right dislocations: R

Any constituent boundary appearing after RS2 has to be labelled with RS1 instead of IC2 because once a right dislocation is opened, all following constituents become additions to

the dislocation.

2.7 Embedded sentences: E

These are all sentences embedded in a matrix sentence that continues after the embedded sentence. In contrast to the former strategy, even very short parentheses (Exdeu glaub ich) are annotated with EM3. If necessary, these short parentheses (less or equal two words) can be relabelled automatically.

2.8 Boundaries at presentential and postsentential discourse particles: T/D

In contrast to the former strategy, we use PM3, if a discourse particle unequivocally can be classified as a confirmation, as in

A: *Paßt Ihnen drei Uhr* SM3
B: *Ja* PM3 *Dann zum zweiten Termin ...*

Much more common is, however, that the particle is followed by a sort of equivalent confirmation, e.g.:

B: *Ja* DS1/TM2 *paßt ausgezeichnet* SM3 *Dann zum zweiten Termin* PM2 *...*

Here, we simply cannot tell apart the two functions 'confirmation' or 'discourse particle'. This is, however, not necessary because in these cases, the functional load on this particle is rather low. It might thus be the most appropriate solution **not** to decide on the one or the other reading but to treat this distinction as neutralized. This means for the higher linguistic modules that, in constellations like this, these particles might simply be treated as discourse particles without any pronounced semantic function; i.e., in the short run, they can be neglected.

Note that presentential particles (at the beginning of a sentence/phrase) and postsentential particles (tags at the end of a sentence/phrase) are annotated with the same label. The could be told apart, however, if one looks at the word boundary of this particle: no M boundary in presentential position, but a M boundary in postsentential position.

2.9 Ambiguous boundaries: A

AM3 and AM2 are ambiguous boundaries between clauses and phrases, resp., and are discussed in more detail in [**?**]. Particles that are very often surrounded by the AC1 label are:

auch: *da müßten doch wohl einige Sachen* AC1 *auch* AC1 *zu Hause aufbereitet werden*
doch: *entscheiden wir uns* AC1 *doch* AC1 *für den ersten Advent*
gleich: *dann halten wir das* AC1 *gleich* AC1 *als ersten Termin fest*
noch: *wollen wir dann* AC1 *noch* AC1 *'n Nachtreffen machen*
schon: *sagen wir dann* AC1 *schon* AC1 *um fünfzehn Uhr*
sogar: *dann würd' es am neunzehnten* AC1 *sogar* AC1 *bei mir funktionieren*
vielleicht: *am einundzwanzigsten* AC1 *vielleicht* AC1 *besser erst elf Uhr*
wieder: *wie sieht es denn aus* AC1 *wieder* AC1 *an einem Mittwoch*

For the automatic assignment of accent position, these particles are treated in a special way as well: the are labelled as A3U, i.e., it cannot be decided with the language model whether they are accented or not.

2.10 Internal constituent boundaries: I

The decision whether to put in an IC2 label or not is very often difficult to take. The criteria for putting in an IC2 label are that (1) the boundary should be really inside the clause, i.e. far from left and right edges, and that (2) the constituent that precedes the boundary is 'prosodically heavy', i.e. normally a NP that can be the carrier of a primary accent.

Cases that are quite clear are:

(a) clauses having no or only one NP between the verbal braces and containing no further particles (usually very short sentences); no IC2 boundary is possible: *das ist gut* or *wir wollen (*IC0*) eine gemeinsame Reise (*IC0*) machen*

(b) If a clause contains two (or more) NPs between its verbal braces, an IC2 boundary often appears between them. In these cases, each NP provides a new, separate piece of information, which makes the phrases prosodically heavy:

... daß wir AM2 *noch* AM2 *im Juni* IC2 *einen Besuch* IC0 *abstatten wollten*

um noch mal 'n Arbeitstreffen IC2 *unter der Woche* IC2 *beim Kaffee abzusprechen*

The problematic clauses are very often those containing a considerable amount of words that are not part of an NP. Such words are usually adverbs or modal particles. They often do not carry any stress originally, but the amount of little words put into the space between the verbal braces finally requires that a rhythmical break has to be be made somewhere in the sentence. However, it is difficult to decide (a) if this break is actually made by all speakers, and, if it is made, (b) where exactly it is made:

da sollten wir vielleicht doch lieber IC2 *vom Montag bis Montag fahren*

Although in this case the IC2 label marks no clear division between two NPs, there is a prosodically marked boundary between *'lieber'* and *'vom'*. Here, the word *'doch'* has probably enough stress to 'play the role' of an NP and make the sentence fall into two

sections, as far as prosody is concerned.

A similar case is: *im März hab' ich eigentlich durchgängig* IC2 *immer irgendwas dazwischen* In this example, it is the word 'durchgängig' that assumes a certain degree of stress and thus makes a prosodic division of the clause into two parts possible.

These phenomena of stress, intonation and rhythm are the main criteria in the IC2 labelling of problematic sentences.

3 Correspondences of M with B and D

Tables 4 and 5 display the correspondences of M with B and M with D: B3: strong, B: weak, B9: irregular, B9: every other boundary; D3: dialog act boundary, D0: no dialog act boundary. These figures are obtained for a subsample of the whole data base comprising 30 dialogs, that is annotated with B labels and with D labels as well. These figures can be taken as an estimation of the frequency of the labels across the whole database that we want to label (CD–ROMs 1, 2–5, 7, 12, 14). For each of the M subclasses (type) it can be seen that the feature 'strength' correlates with the prosodic–perceptual marking: the higher the strength, the higher the number of B3 or B2 labels. This overall tendency holds for the correspondence with the D labels as well.

4 M labels for English data bases

Most of the German M labels can be used for the English data as well. Some labels, however, were redefined, and some new labels were introduced. Tables 7 and 6 displays those M labels that were used for English, together with one exan'mple each.

For three German labels, RS2, RC1, and TM2, no corresponding defining context could be found in English: for RS2 and RC1, because there is no verbal brace in English, and for TM2, because the English transliteration does not reliably annotate pauses etc.

Up to mow, some ten English dialogs are labelled. It can be seen in Tables 7 and 6 that for four labels, PC2, LS2, LC2, and DS3, no tokens were found in these dialogs. If this holds across all English dialogs or if there are only a few of them, these labels can be discarded or merged with related labels.

For all these reasons, the English labels are thus not fixed yet but can change according to our experiences with the annotations to come. The following labels are introduced especially for the English data:

SM3E, SS3E:
A progressive form is not impossible in German, but sounds rather pretentious and obsolete; as it is rather common in English, we label it in a special way. If necessary, these two labels

	#	B3	B2	B9	B0
SM3	654	82.72	8.87	3.06	5.35
SM2	154	57.79	25.97	0.65	15.58
SS2	25	76.00	8.00	0.00	16.00
SM1	46	28.26	30.43	0.00	41.30
SS1	3	0.00	0.00	0.00	100.00
SC3	24	70.83	29.17	0.00	0.00
SC2	19	52.63	36.84	0.00	10.53
PM3	235	76.60	12.34	2.98	8.09
PC2	32	59.38	12.50	0.00	28.12
PM1	23	4.35	8.70	0.00	86.96
LS2	45	53.33	28.89	0.00	17.78
LC2	27	37.04	18.52	0.00	44.44
RS2	98	54.08	18.37	2.04	25.51
RC2	67	37.31	19.40	1.49	41.79
RC1	84	29.76	15.48	1.19	53.57
EM3	44	29.55	34.09	4.55	31.82
TM2	40	65.00	25.00	7.50	2.50
DS3	22	54.55	27.27	0.00	18.18
DS1	512	19.14	34.38	2.93	43.55
AM3	189	43.39	15.34	6.88	34.39
AM2	23	21.74	13.04	4.35	60.87
AC1	348	4.89	9.48	3.16	82.47
IC2	367	23.98	24.52	7.08	44.41
IC1	16	31.25	25.00	0.00	43.75
IC0	10182	1.03	2.93	5.08	90.96

Table 5: Correspondence of M labels with B labels

	D3	D0
SM3	92.70	7.30
SM2	25.00	75.00
SS2	40.91	59.09
SM1	4.76	95.24
SS1	0.00	100.00
SC3	52.94	47.06
SC2	33.33	66.67
PM3	45.75	54.25
PC2	0.00	100.00
PM1	4.17	95.83
LS2	9.52	90.48
LC2	3.85	96.15
RS2	10.34	89.66
RC2	0.00	100.00
RC1	0.00	100.00
EM3	8.33	91.67
TM2	18.52	81.48
DS3	52.94	47.06
DS1	6.65	93.35
AM3	34.48	65.52
AM2	4.17	95.83
AC1	0.31	99.69
IC2	0.30	99.70
IC1	8.33	91.67
IC0	0.70	99.30

Table 6: Correspondence of M labels with D labels

main class	label	context (between/at) with example
		sentences
		Main clause and main clause:
M3	SM3	see you then SM3 have a nice seminar
		Main clause and subordinate clause:
M3	SM2	what would be a good time SM2 to meet again
		Subordinate clause and main clause:
M3	SS2	you are out through June second SS2 did you say
		main clause and subord. clause (progr. form):
M3	SM2E	I will send you mail SM2E regarding the location
		subord. (progr. form) and main clause:
M3	SS2E	Looking at my schedule SS2E I am free in the afternoon
		Main clause and subordinate clause, prosodically integrated:
M3	SM1	I guess SM1 we should try and get together
		Subordinate clause and main clause, prosodically integrated:
M3	SS1	two to three hours SS1 you say
		Coordination of main clauses and of subordinate clauses:
M3	SC3	maybe we can get together SC3 and discuss the planning
		Subordinate clause and subordinate clause:
M3	SC2	you will have an extra week to do all the stuff SC2 that you wanted
		free Phrases:
		free Phrase, stand alone:
M3	PM3	thanks PM3 bye
		sequence in free Phrases:
M2	PC2	two to four p.m. PC2 on Saturday PC2 the second PC2 of October
		free Phrase, prosodically integrated, no dialogue act boundary:
M3	PM1	then Friday one o'clock PM1 two hours
		Left dislocations:
		Left dislocation:
M3	LS2	on the eighth LS2 that would be good
		left dislocation (without anaphor. ref.):
M3	LS2E	Wednesday through Friday LS2E I am like in seminars
		sequence of Left dislocations:
M2	LC2	sometime LC2 in the afternoon LC2 like maybe two LC2 'till five LS2 how is then
		seq. of left dislocations (without anaphor. ref.):
M2	LC2E	most days LC2E Monday through Friday I have classes
		conjunction at beginning of clause (without *and, or*):
M2/M0	CS1E	because CS1 Wednesday through Friday ...
		Right dislocations:
		Right dislocation (any part after completion):
M3	RS2E	to think up something really nice RS2E for him
		sequence of Right dislocations:
M2	RC2	I am free on Monday RC2 except forten to twelve RC2 in the morning
		possible right 'dislocation' (different meaning):
M2	RC1E	I am free RC1E on Monday

Table 7: Examples for English boundary labels and their context, part I.

main class	label	context (between/at) with example
Embedded strings:		
M3	EM3	**Embedded sentence/phrase:** we will have to EM3 you know EM3 look for something ...
Discourse particles:		
MU	DS3	**pre-/postsentential particle, ambisentential:** NO DATA
MU	DS1E	**pre-/postsentential particle:** well DS1 I am out of town ...
Ambiguous boundaries:		
MU	AM3	**between sentences, Ambiguous:** I am out of town AM3 the thirtieth through the third AM3 I am in San Francisco
MU	AM2	**between free phrases, Ambiguous:** okay AM2 then AM2 Friday one o'clock
MU	AC1	**between constituents, Ambiguous:** ... and then AC1 maybe AC1 in the meantime ...
Constituents:		
M2	IC2	**between Constituents:** what would be IC2 a good time
M2	IC1	**asyndetic listing of Constituents:** how 'bout the twenty seventh IC1 twenty eighth or thirty first
default, no boundary:		
M0	IC0	**every other word boundary:** if we cannot make it

Table 8: Examples for English boundary labels and their context, part II.

can be merged with SM3 and SS3.

LS2E, LC2E :
In these left dislocations or sequences of left dislocations, no anaphoric reference can be found in the corresponding matrix sentence.

CS1:
This label denotes conjunctions at the beginning of sentences, but not *and* or *or*.

RS2E:
This label denotes any part of sentences after their grammatical completion; it replaces German RS2 and RC1 because their defining criterion (end of verbal brace) does not exist in English.

RC1E:
This is a boundary at a possible right dislocation, but its omission leads to a different meaning of the utterance.

DS1E:
This label is used as well for particles followed by a pause that in German are labelled differently with TM2 because both types cannot be told apart in the English transliteration.

DS3:

This label has not been annotated in our material; possibly, such particles do not occur in ambisentential positions in English.

5 Concluding remarks