A Framework for Annotating Information Structure in Discourse

Sasha Calhoun[1], **Malvina Nissim**[1], **Mark Steedman**[1] **and Jason Brenier**[2]

[1]Institute for Communicating and Collaborative Systems, University of Edinburgh, UK
Sasha.Calhoun@ed.ac.uk, {steedman,mnissim}@inf.ed.ac.uk
[2]Department of Linguistics, University of Colorado at Boulder
jbrenier@colorado.edu

Abstract

We present a framework for the integrated analysis of the textual and prosodic characteristics of information structure in the *Switchboard* corpus of conversational English. Information structure describes the availability, organisation and salience of entities in a discourse model. We present standards for the annotation of *information status* (old, mediated and new), and give guidelines for annotating *information structure*, i.e. *theme/rheme* and *background/kontrast*. We show that information structure in English can only be analysed concurrently with prosodic prominence and phrasing. This annotation, using stand-off XML in NXT, can help establish standards for the annotation of information structure in discourse.

1 Introduction

We present a framework for the integrated analysis of the textual and prosodic characteristics of information structure in a corpus of conversational English. Section 2 introduces the corpus as well as the tools we employ in the annotation process. We propose two complementary annotation efforts within this framework. The first, information status (*old, mediated, new*), expresses the *availability* of entities in discourse (Section 3). The second scheme will firstly annotate *theme/rheme*, i.e. how each intonation phrase is organised in the discourse model, and secondly *kontrast*: how *salient* the speaker wishes to make each entity, property or relation (Section 4). We will demonstrate that the perception of both of these is intimately affected by prosodic structure. In particular, the theme/rheme division affects prosodic phrasing; and information status and kontrast affect relative prosodic prominence. Therefore we also propose to annotate a subset of the corpus for this prosodic information (Section 5). In conjunction with existing annotations of the corpus, our integrated framework using NXT will be unique in the field of conversational speech in terms of size and richness of annotation.

2 Corpus and Tools

The Switchboard Corpus (Godfrey et al., 1992) consists of 2430 spontaneous phone conversations (average six minutes), between speakers of American English, for three million words. The corpus is distributed as stereo speech signals with an orthographic transcription per channel time-stamped at the word level. A third of this is syntactically parsed as part of the Penn Treebank (Marcus et al., 1993) and has dialog act annotation (Shriberg et al., 1998). We used a subset of this. In adherence with current standards, we converted all the existing annotations, and are producing the new discourse annotations in a coherent multi-layered XML-conformant schema, using NXT technology (Carletta et al., 2004).[1] This allows us to search over and integrate information from the many layers of annotation, including the

[1]Beside the NXT tools, we also used the TIGER Switchboard filter (Mengel and Lezius, 2000) for the XML-conversion. Using existing markup we automatically selected and filtered NPs to be annotated, excluding locative, directional, and adverbial NPs and disfluencies, and adding possessive pronouns. See (Nissim et al., 2004) for technical details.

Proceedings of the Workshop on Frontiers in Corpus Annotation II: Pie in the Sky, pages 45–52,
Ann Arbor, June 2005.

sound files. NXT tools can be easily customised to accommodate different layers of annotation users want to add, including data sets that have low-level annotations time-stamped against a set of synchronized signals, multiple, crossing tree structures, and connection to external corpus resources such as gesture ontologies and lexicons (Carletta et al., 2004).

3 Information Status

Information Status describes how *available* an entity is in the discourse. We define this in terms of the speaker's assumptions about the hearer's knowledge/beliefs, and we express it by the well-known old/new distinction.[2]

3.1 Annotation Scheme

Our annotation scheme for the discourse layer mainly builds on (Prince, 1992) and (Eckert and Strube, 2001), as well as on related work on annotation of anaphoric links (Passonneau, 1996; Hirschman and Chinchor, 1997; Davies et al., 1998; Poesio, 2000). Prince defines "old" and "new" with respect to the *discourse model* as well as the *hearer*'s point of view. Considering the interaction of both these aspects, we define as *new* an entity which has not been previously referred to and is yet unknown to the hearer, and as *mediated* an entity that is newly mentioned in the dialogue but that the hearer can *infer* from the prior context.[3] This is mainly the case of generally known entities (such as "the sun", or "the Pope" (Löbner, 1985)), and *bridging* (Clark, 1975), where an entity is related to a previously introduced one. Whenever an entity is not new nor mediated is considered as *old*.

Because finer-grained distinctions (e.g. (Prince, 1981; Lambrecht, 1994)) have proved hard to distinguish reliably in practice, we organise our scheme *hierarchically*: we use the three main classes described above as top level categories for which more specific subtypes can assigned. This approach preserves a high-level, more reliable distinction while allowing a finer-grained classification that can be exploited for specific tasks.

Besides the main categories, we introduce two more classes. A category non-applicable is used for wrongly extracted markables (such as "course" in "of course"), for idiomatic occurrences, and expletive uses of "it". Traces are automatically extracted as markables, but are left unannotated. In the rare event the annotators find some fragments too difficult to understand, a category not-understood can be assigned. Entities marked as non-applicable or not-understood are excluded from any further annotation. For all other markables, the annotators must choose between old, mediated, and new. For the first two, subtypes *can* also be specified: subtype assignment is encouraged but not compulsory.

New The category new is assigned to entities that have not yet been introduced in the dialogue and that the hearer cannot infer from previously mentioned entities. No subtypes are specified for this category.

Mediated Mediated entities are inferrable from previously mentioned ones, or generally known to the hearer. We specify nine subtypes: general, bound, part, situation, event, set, poss, func_value, aggregation.[4] Generally known entities such as "the moon" or "Italy" are assigned a subtype general. Most proper nouns fall into this subclass, but the annotator could opt for a different tag, depending on the context. Also mediated are bound pronouns, such as "them" in (1), which are assigned a subtype bound.[5]

(1) [...] it's hard to raise *one child* without **them** thinking they're the pivot point of the universe.

A subtype poss is used to mark all kinds of intra-phrasal possessive relations (pre- and postnominal).

Four subtypes (part, situation, event, and set) are used to mark instances of bridging. The subtype part is used to mark part-whole relations for physical objects, both as intra- and inter-phrasal relations. (This category is to be preferred to poss whenever applicable.) The occurrence of "the door" in (2), for instance, is annotated as mediated/part.

(2) When I come *home* in the evenings my dog greets me at **the door**.

For similar relations that do not involve physical objects, i.e. if an entity is part of a situation set up by

[2] We follow Prince in using "old" rather than "given" to refer to "not-new" information, but regard the two as identical.

[3] This type corresponds to Prince's (1981; 1992) *inferrables*.

[4] Some of the subtypes are inspired by categories developed for bridging markup (Passonneau, 1996; Davies et al., 1998).

[5] All examples in this paper are from the Switchboard Corpus. The markable in question is typed in boldface; antecedents or trigger entities, where present, are in italics. For the sake of space we do not provide examples for each category (see (Nissim, 2003)).

a previously introduced entity, we use the subtype situation.[6],as for the NP "the specifications" in (3).

(3) I guess I don't really have a problem with *capital punishment*. I'm not really sure what **the exact specifications** are for Texas.

The subtype event is applied whenever an entity is related to a previously mentioned verb phrase (VP). In (4), e.g., "the bus" is triggered by *travelling around Yucatan*.

(4) We were *travelling around Yucatan*, and **the bus** was really full.

Whenever an entity referred to is a subset of, a superset of, or a member of the same set as a previously mentioned entity, the subtype set is applied.

Rarely, an entity refers to a value of a previously mentioned function, as "zero" and "ten" in (5). In such cases a subtype func-value is assigned.

(5) I had kind of gotten used to *centigrade temperature* [...] if it's between **zero** and **ten** it's cold.

Lastly, a subtype aggregation is used to classify coordinated NPs. Two old or med entities, for instance do not give rise to an old coordinated NP, unless it has been previously introduced as such. A mediated/aggregation tag is assigned instead.

Old An entity is old when it is not new nor mediated. This is usually the case if an entity is *coreferential* with an already introduced entity, if it is a generic pronoun, or if it is a personal pronoun referring to the dialogue participants. Six different subtypes are available for old entities: identity, event, general, generic, ident_generic, relative. In (6), for instance, "us" would be marked as old because it corefers with "we", and a subtype identity would also be assigned.

(6) [...] *we* camped in a tent, and uh there were two other couples with **us**.

In addition, a coreference link is marked up between anaphor and antecedent, thus creating anaphoric chains (see also (Carletta et al., 2004)). The subtype event applies whenever the antecedent is a VP. In (7), "it" is old/event, as its antecedent is the VP "educate three". As we do not extract VPs as markables, no link can be marked up.

(7) I most certainly couldn't *educate three*. I don't know how my parents did **it**.

[6]This includes elements of the thematic grid of an already introduced entity. It subsumes Passonneau's (1996) class "arg".

Also classified as old are personal pronouns referring to the dialogue participants as well as generic pronouns. In the first case, a subtype general is specified, whereas the subtype for the second is generic. An instance of old/generic is "you" in (8).

(8) up here **you** got to wait until Aug- August until the water warms up.

In a chain of generic references, the subtype ident_generic is assigned, and a coreference link is marked up. Coreference is also marked up for relative pronouns: they receive a subtype relative and are linked back to their head.

The guidelines contain a decision tree the annotators use to establish priority in case more than one class is appropriate for a given entity. For example, if a mediated/general entity is also old/identity the latter is to be preferred to the former. Similar precedence relations hold among subtypes.

To provide more robust and reliable clues in annotating bridging types (e.g. for distinguishing between poss and part), we provided replacement tests and referred to relations encoded in knowledge bases such as WordNet (Fellbaum, 1998) (for part) and FrameNet (Baker et al., 1998) (for situation).

3.2 Validation of the Scheme

Three Switchboard dialogues (for a total of 1738 markables) were marked up by two different annotators for assessing the validity of the scheme. We evaluated annotation reliability by using the Kappa statistic (Carletta, 1996). Good quality annotation of discourse phenomena normally yields a kappa (K) of about .80. We assessed the validity of the scheme on the four-way classification into the three main categories (old, mediated and new) and the non-applicable category. We also evaluated the annotation including the subtypes. All cases where at least one annotator assigned a not-understood tag were excluded from the agreement evaluation (14 markables). Also excluded were all traces (222 markables), which the annotators left unmarked. The total markables considered for evaluation over the three dialogues was therefore 1502.

The annotation of the three dialogues yielded $K = .845$ for the high-level categories, and $K = .788$ when including subtypes ($N = 1502$; $k = 2$).[7]

[7]N stands for the number of instances annotated and k for

These results show that overall the annotation is reliable and that therefore the scheme has good reproducibility. When including subtypes agreement decreases, but backing-off to the high-level categories is always possible, thus showing the virtues of a hierarchically organised scheme. Reliability tests for single categories showed that mediated and new are more difficult to apply than old, for which agreement was measured at $K = .902$, although still quite reliable ($K = .800$ and $K = .794$, respectively). Agreement for non-applicable was $K = .846$.

The annotators found the decision tree very useful when having to choose between more than one applicable subtype, and we believe it has a significant impact on the reliability of the scheme.

The scheme was then applied for the annotation of a total of 147 Switchboard dialogues. This amounts to 43358 sentences with 69004 annotated markables, 35299 of which are old, 23816 mediated and 9889 new (8127 were excluded as non-applicable, and 160 were not understood), and 16324 coreference links.

In Section 6 we use this scheme to annotate the Pie-in-the-Sky text.

3.3 Related Work

To our knowledge, (Eckert and Strube, 2001) is the only other work that explicitly refers to IS annotation. They also use a Prince's (1992)-based old/med/new distinction for annotating Switchboard dialogues. However, their IS annotation is specifically designed for salience ranking of candidate antecedents for anaphora resolution, and not described in detail. They do not report figures on inter-annotator agreement so that a proper comparison with our experiment is not feasible. Among the schemes that deal with annotation of anaphoric NPs, our scheme is especially comparable with DRAMA (Passonneau, 1996) and MATE (Davies et al., 1998). Both schemes have a hierarchical structure. In DRAMA, types of *inferrables* can be specified, within a division into conceptual (pragmatically determined) vs. linguistic (based on argument structure) inference. No annotation experiment with inter-annotator agreement figures is however reported. MATE provides subtypes for bridging relations, but they were not applied in any annotation excercise, so that reliability and distribution of categories are only based on the "core scheme" (true coreference). For a detailed comparison of our approach with related efforts on the annotation of anaphoric relations, see (Nissim et al., 2004).

the number of annotators. Unless otherwise specified, $N = 1502$ and $k = 2$ hold for all K scores reported in Section 3.

4 Information Structure

We have seen that information status describes how available an entity is in a discourse. Generally *old* entities are available, and *new* entities are not. In prosody we find that newness is highly correlated with pitch accenting, and oldness with deaccenting (Cutler et al., 1997). However, this is only one aspect of information structure. We also need to describe how speakers signal the organisation and salience of elements in discourse. Building on the work of (Vallduví & Vilkuna, 1998), as developed by (Steedman, 2000), we define two notions, *theme/rheme* structure and *background/kontrast*.

Theme/rheme structure guides how an element fits into the discourse model: if it relates back it is *thematic*; if it advances the discourse it is *rhematic*. Steedman claims that intonational phrases can mark information units (*theme* and *rheme* - though not all boundaries are realised and a unit may contain more than one phrase). The pitch contour associated with nuclear accents in themes is distinct from that in rhemes (which he identifies as L+H*LH% and H*LH% re ToBI (Beckman and Elam, 1997)), so that, where present, such boundaries disambiguate information structure. (See (9)).[8]

(9) (Q) Personally, I love hyacinths.
What kind of bulbs grow well in your area?
(A)

(In	MY	AREA)	
Bkgd	*Kont.*	*Bkgd*	*(Theme)*
(it is the	DAFFODIL)		
Bkgd	*Kont.*		*(Rheme)*

The second dimension, *kontrast*, relates to salience.[9] We expect new entities to be salient and old entities not. Therefore, if an old element is salient, or a new one especially salient, an extra meaning is implied.

[8] Annotation is as in Section 3. Words in SMALL CAPS are accented, parentheses indicate intonation phrases, including boundary tones if present. See website to hear some examples from this section.

[9] We use *kontrast* to distinguish it from the everyday use of *contrast* and the sometimes conflicting uses of *contrast* in the literature. Annotators, however, will not be given this term.

These are largely subsumed by *kontrast*, i.e. distinguishing an element from alternatives made available by the context (See (9)).

4.1 Annotation Scheme

As we have seen, in English, information structure is primarily conveyed by intonation. We therefore think it is vital for annotators to listen to the speech while annotating this structure.

4.1.1 *Theme/Rheme*

We have claimed that prosodic phrasing can divide utterances into information units. However, often theme material is entirely background, i.e., mutually known and without contrasting alternatives. Therefore, for both model theoretic and practical purposes, it is the same as background of the rheme. Accordingly, we work with a test for themehood, defining the **rheme** as any prosodic phrase that is not identifiable as a **theme**.

Annotators will mark each prosodic phrase as a *theme* if it only contains information which links the utterance to the preceding context, i.e. setting up what they're saying in relation to what's been said before. In their opinion, even if this is not the tune the speaker used, it must sound appropriate if they say it with a highly marked tune, such as L+H* LH%. For example, in (10), the phrase "where I lived" links "was a town called Newmarket" to the statement the speaker lived in England (accenting not shown). It would be appropriate to utter it with an L+H* accent on "Where" and/or "lived,", and a final LH%. So it is a theme. The same accent on "town" and/or "Newmarket" sounds inappropriate, and it advances the discussion, so it is a rheme.

(10) I lived over in England for four years
(Where I lived) *(Theme)*
(was a town called Newmarket) *(Rheme)*

4.1.2 *Background/Kontrast*

Although there is a clear link between prosodic prominence and kontrast, there are a number of disagreements about how this works which this annotation effort seeks to resolve. Some, including (Steedman, 2000), have claimed that kontrast within theme and kontrast within rheme are marked by categorically distinct pitch accents. Another view is that kontrast, also called contrastive focus or topic, only applies to *themes* that are contrastive; the head of a rheme phrase always attracts a pitch accent, it is therefore redundant to call one part kontrastive. Further, some consider kontrast within a rheme phrase only occurs when there is a clear alternative set, i.e. the distinction between broad and narrow focus, as in (9) where *daffodil* contrasts with other bulbs the speaker might grow. Again, there is controversy on whether there is an intonational difference between broad and narrow focus (Calhoun, 2004a). If these distinctions are marked prosodically, it is disputed whether this is with different pitch accents (Steedman), or by the relative height of different accents in a phrase (Rump and Collier, 1996; Calhoun, 2004b).

Rather than using the abstract notion of kontrast directly, annotators will identify discourse scenarios which commonly invoke kontrast (drawing on functions of emphatic accents from (Brenier et al., 2005)).[10] This addresses the disagreements above, while making our annotation more constrained and robust. In each case, using the full discourse context including the speech, annotators mark each content word (noun, verb, adjective, adverb and demonstrative pronoun) for the first category that applies. If none apply, they mark it as **background**.

correction The speaker's intent is to correct or clarify another just used by them or the other speaker. In (11), e.g., the speaker wishes to clarify whether her interlocutor really meant "hyacinths".

(11) (now are you sure they're **HYACINTHS**) (because that is a BULB)

contrastive The speaker intends to contrast the word with a previous one which was (a) a current topic; (b) semantically related to the contrastive word, such that they belong to a natural set. In (12), B contrasts recycling in her town "San Antonio", with A's town "Garland", from the set *places where the speakers live*.

(12) (A) I live in *Garland*, and we're just beginning to build a real big recycling center...
(B) (YEAH there's been) (NO emphasis on recycling at ALL) (in **San ANTONIO**)

[10] Emphasis can occur for two major reasons, both identified by Brenier: emphasis of a particular word or phrase, i.e. kontrast, or emphasis over a larger span of speech, conveying affective connotations such as excitement, which is not included here. (Ladd, 1996).

subset The speaker highlights one member of a more general set that has been mentioned and is a current topic. In (13), the speaker introduces "three day cares", and then gives a fact about each.

(13) (THIS woman owns *THREE day cares*) (**TWO** in Lewisville) (and **ONE** in Irving) (and she had to open **the SECOND one** up) (because her WAITING list was) (a YEAR long)

adverbial The speaker uses a focus-sensitive adverb, i.e. *only*, *even*, *always* or *especially* to highlight that word, and not another in the natural set. The adverb and/or the word can be marked. In (14), B didn't even like the "previews" of 'The Hard Way', let alone the movie.

(14) (A) I like Michael J Fox, though I thought he was crummy in 'The Hard Way'.
(B) (I didn't even like) (the **PREVIEWS**)

answer The word (or its syntactic phrase, e.g. an NP) and no other, fills to an open proposition set up in the context. It must make sense if they had only said that word or phrase. In (15), A sets up the "blooms" she can't identify, and B answers "lily".

(15) (A) We have *these blooms*, I'm not sure what they are but they come in all different colours yellow, purple, white...
(B) (I BET you) (that that's a **LILY**)

Again, in Section 6 we apply the scheme to the Pie-in-the-Sky text.

4.2 Related Work

Annotator agreement for pitch accents and prosodic boundaries, re ToBI, is about 80% and 90% respectively (Pitrelli et al., 1994). Automatic performance, using acoustic and textual features, is now above 85% accuracy (Shriberg et al., 2000). However, this does not distinguish prosodic events which occur for structural or rhythmical reasons from those which mark information structure (Ladd, 1996). (Heldner et al., 1999) try to predict focal accents. They define this minimally as the most prominent in a three-word phrase. (Hirschberg, 1993) got 80-98% accuracy using only text-based features. However, her definition of contrast was not as thorough as ours. (Hedberg and Sosa, 2001) looked at marking of ratified, unratified (old and new) and contrastive topics and foci (theme and rheme) with ToBI pitch accents. (Baumann et al., 2004) annotated a simpler information structure and prosodic events in a small German corpus.

5 Information Structure and Prosodic Structure

Much previous work, not corpus-based, draws a direct correspondence between information structure, prosodic phrasing and pitch accent type. However in real speech there are many non-semantic influences on prosody, including phrase length, speaking rate and rhythm. Information structure is rather a strong constraint on the realisation of prosodic structure (Calhoun, 2004a). Contrary to the assumption of ToBI, this structure is metrical, highly structured and linguistically relevant both within and across prosodic phrases (Ladd, 1996; Truckenbrodt, 2002).

One of our main aims is to test how such evidence can be reconciled with theories presented earlier about the relationship between information structure and prosody. Local prominence levels have been shown to aid in the disambiguation of focal adverbs, anaphoric links, and global discourse structures marked as *elaboration*, *continuation*, and *contrast* (Dogil et al., 1997). Global measures of prominence level have been linked to topic structure, corrections, and turn-taking cues (Ayers, 1994). (Brenier et al., 2005) found that *emphatic* accents realised special discourse functions such as *assessment*, *clarification*, *contrast*, *negation* and *protest* in child-directed speech. Most of these functions can be seen as conversational implicatures of kontrast, i.e. if an element is unexpectedly highlighted, this implies an added meaning. Brenier found that while pitch accents can be detected using both acoustic and textual cues; textual features are not useful in detecting emphatic pitch accents, showing there is added meaning not available from the text.

As noted in Section (4.2), inter-annotator agreement for the identification of prosodic phrase boundaries with ToBI is reasonably good. We will therefore label ToBI break indices 3 and 4 (conflated) (Beckman and Elam, 1997). Annotators will also mark the perceived level of prosodic prominence on each word using a defined scale. We are currently running a pilot experiment to identify a reasonable number of gradations of prosodic prominence, from completely unstressed and/or reduced to highly emphatic, to use for the final annotation.

[But [[[Yemen' s]$_{med/general}$ president]$_{med/poss}$]]$_{Contrastive}$ says]$_{THEME}$ [[the FBI]$_{old/identity}$ has told [him]$_{old/identity}$]$_{THEME}$ [[the explosive material]$_{med/set}$ could only have come from [[[the U.S.]$_{med/general}$, [israel]$_{med/general}$, or [[two arab countries]$_{med/set}$]$_{med/aggregation}$.]$_{Adverbial}$]$_{RHEME}$ [And to [[a former federal bomb investigator]$_{new}$,]$_{Contrastive}$]$_{THEME}$ [[that description]$_{old/event}$ suggests]$_{THEME}$ [[a powerful military-style plastic explosive C-4]$_{med/set}$]$_{Answer}$ [[that]$_{old/relative}$ can be cut or molded into [different shapes]$_{new}$.]$_{RHEME}$

Figure 1: Annotation of Pie-in-the-Sky sentences with Information Structure

6 Pie-in-the-Sky annotation

"Pie in the Sky" is a joint effort to annotate two sentences with as much semantic/pragmatic information as possible (see `http://nlp.cs.nyu.edu/meyers/pie-in-the-sky.html`). Information structure is one of the desired annotation layers. And, as standards are not yet established, our proposal contributes to defining annotation guidelines for this structure. Figure 1 report the Pie-in-the-sky sentences enriched with our annotation. The context prior to these sentences is as follows:

> "a 12-year-old boy reports seeing a man launch a rubber boat from a car parked at the harbor. fbi officials find what they believe may be explosives in the car. yemeni police trace the car to a nearby house. the fbi finds traces of explosives on clothes found neighbors say they saw two men who they describe as "arab-looking" living there for several weeks. police also find a second house where authorities believe two others may have assembled the bomb, possibly doing some welding. passports found in one of the houses identify the men as from a privilege convenience province noted for lawless tribes. but the documents turn out to be fakes. meantime, analysts at the fbi crime lab try to discover what the bomb was made from. no conclusions yet, u.s. officials say. but a working theory, plastic explosive."

We identified 14 NPs markable for information status (see Figure 1).[11] Most annotations were straightforward. Some comments though: "Yemen" is annotated as med/general, although it could also be med/sit as "Yemeni" was previously mentioned. Our decision tree was used for such cases. "The explosive material" is med/set not old/identity since it refers to the kind of explosive used rather than to a specific entity previously mentioned.

In the absence of any prosodic annotation in the transcript, these sentences are slightly ambiguous as to information structure. The most likely interpretation is given in Figure 1.[12] For example, "Yemen's President" contrasts with "US officials", in the set of people talking about what the bomb is made of. Since both words are contrastive, either or both could have L+H* accents, whereas "say" could not. The inclusion of the latter in the theme is consistent with the possibility of a rising boundary LH% after it. "The FBI has told him" is thematic because it links "Yemen's president"'s opinion to the previous discourse. It also would sound appropriate with an L+H*LH% tune. As can be seen, although theme/rheme and prosodic phrase boundaries align, in both cases the VP is split between information/intonation phrases. The independence of information structure and intonation structure from traditional surface structure is a major reason behind our use of 'stand-off' markup.

[11] Square brackets are used to mark annotation boundaries.

[12] Kontrast is marked with the relevant category, unmarked words are background.

7 Applications and Future Work

Once completed, the annotations we have presented, along with those existing for syntax, disfluencies and dialog-acts on the same portion of *Switchboard*, will create a corpus of conversational speech unique in terms of size and richness of annotation. In conjunction with the NXT tools, this resource would optimally lend itself to detailed and rich analysis of diverse linguistic phenomena, the ultimate goal of the Pie in the Sky project. It will be useful for a large range of NLP applications, including paraphrase analysis and generation, topic detection, information extraction and speech synthesis in dialogue systems.

Website Example sound files available at *http://homepages.inf.ed.ac.uk/s0199920/pieinsky.html*.

Acknowledgements Part of this work was funded by Scottish Enterprise (The Edinburgh-Stanford Link *Paraphrase Analysis for Improved Generation* and *Sounds of Discourse*). We would like to thank David Beaver, Jean Carletta, Shipra Dingare, Florian Jaeger, Dan Jurafsky, Vasilis Karaiskos and Bob Ladd for valuable help and discussion.

References

G. M. Ayers. 1994. Discourse functions of pitch range in spontaneous and read speech. In J. Venditti, editor, *OSU Working Papers in Linguistics*, volume 44, pages 1–49.

C. F. Baker, C. J. Fillmore, and J. B. Lowe. 1998. The Berkeley FrameNet project. In C. Boitet and P. Whitelock, editors, *Proc. COLING-ACL*, pages 86–90.

E. Bard, D. Robertson, and A. Sorace. 1996. Magnitude estimation of linguistic acceptability. *Language*, 72(1):32–68.

S. Baumann, C. Brinckmann, S. Hansen-Schirra, G-J. Kruijff, I. Kruijff-Korbayová, S. Neumann, and E. Teich. 2004. Multi-dimensional annotation of linguistic corpora for investigating information structure. In *Proc. NAACL/HLT "Frontiers in Corpus Annotation"*, Boston, MA.

M. Beckman and G. Elam. 1997. Guidelines for ToBI Labelling.. The OSU Research Foundation, v.3.0.

P. Boersma and D. Weenink. 2003. Praat:doing phonetics by computer. http://www.praat.org.

J. M. Brenier, D. M. Cer, and D. Jurafsky. 2005. Emphasis detection in speech using acoustic and lexical features. In *LSA Annual Meeting*, Oakland, CA.

S. Calhoun. 2004a. Overloaded ToBI and what to do about it: An argument for function-based phonological intonation categories. In *Univ. of Edinburgh Ling. Postgrad. Conf.*.

S. Calhoun. 2004b. Phonetic dimensions of intonational categories - L+H* and H*. In *Prosody 2004*, Nara, Japan.

J. Carletta, S. Dingare, M. Nissim, and T. Nikitina. 2004. Using the NITE XML Toolkit on the Switchboard Corpus to study syntactic choice: a case study. In *Proc. of LREC2004, Lisbon.*

J. Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *Comp. Ling.*, 22(2):249–254.

H. H. Clark. 1975. Bridging. In R. Schank and B. Nash-Webber, eds, *Theoretical Issues in NLP*. MIT Press, Cambridge, MA.

A. Cutler, D. Dahan, and W. van Donselaar. 1997. Prosody in the comprehension of spoken language: A literature review. *Lang. and Sp.*, 40(2):141–201.

S. Davies, M. Poesio, F. Bruneseaux, and L. Romary. 1998. Annotating coreference in dialogues: Proposal for a scheme for MATE, `http://www.hcrc.ed.ac.uk/~poesio/anno_manual.html`.

G. Dogil, J. Kuhn, J. Mayer, G. Mhler, and S. Rapp. 1997. Prosody and discourse structure: Issues and experiments. In *Proc. of the ESCA Workshop on Intonation: Theory, Models and Applications*, pages 99–102, Athens, Greece.

M. Eckert and M. Strube. 2001. Dialogue acts, synchronising units and anaphora resolution. *J. of Semantics*, 17(1):51–89.

C. Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

J. Godfrey, E. Holliman, and J. McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for research and development. In *Proc. ICASSP-92*, pages 517–520.

N. Hedberg and JM. Sosa. 2001. The prosodic structure of topic and focus in spontaneous english dialogue. In *LSA Workshop on Topic and Focus*, Santa Barbara.

M. Heldner, E. Strangert, and T. Deschamps. 1999. A focus detector using overall intensity and high frequency emphasis. In *Proc. ICPhS-99*, vol 2, 1491–1493, San Francisco.

J. Hirschberg. 1993. Pitch accent in context: Predicting intonational prominence from text. *AI*, 63:305–340.

L. Hirschman and N. Chinchor. 1997. MUC-7 coreference task definition. In *Proc. of 7^{th} Conf. on Message Understanding*.

D. R. Ladd. 1996. *Intonational Phonology*. CUP, UK.

K. Lambrecht. 1994. *Information structure and sentence form. Topic, focus, and the mental representation of discourse referents*. Camb. U. Press, UK.

S. Löbner. 1985. Definites. *J. of Semantics*, 4:279–326.

M. Marcus, B. Santorini, and MA. Marcinkiewicz. 1993. Building a large annotated corpus of english: The Penn treebank. *Comp. Ling.*, 19:313–330.

A. Mengel and W. Lezius. 2000. An XML-based encoding format for syntactically annotated corpora. In *Proc. LREC2000*,121–126.

M. Nissim. 2003. Annotation scheme for information status in dialogue. HCRC, University of Edinburgh. Unpub. ms.

M. Nissim, S. Dingare, J. Carletta, and M. Steedman. 2004. An annotation scheme for information status in dialogue. In *Proc. LREC2004, Lisbon.*

R. Passonneau. 1996. Instructions for applying discourse reference annotation for multiple applications (DRAMA). Unpub. ms..

J. Pitrelli, M. Beckman, and J. Hirschberg. 1994. Evaluation of prosodic transcription labelling reliability in the ToBI framework. In *Proc. of the 3^{rd} Intl. Conf. on Spoken Lge. Proc.*, vol. 2, pages 123–126.

M. Poesio. 2000. The GNOME annotation scheme manual (v.4), `http://www.hcrc.ed.ac.uk/~gnome/anno_manual.html`.

E. F. Prince. 1981. Toward a taxonomy of given-new information. In P. Cole, ed., *Radical Pragmatics*. Acad. Press, NY.

E. Prince. 1992. The ZPG letter: subjects, definiteness, and information-status. In S. Thompson and W. Mann, eds., *Discourse description: diverse analyses of a fund raising text*, pages 295–325. John Benjamins, Philadelphia/Amsterdam.

H.H. Rump and R. Collier. 1996. Focus conditions and the prominence of pitch-accented syllables. *Lang. and Sp.*, 39:1–17.

E. Shriberg, P. Taylor, R. Bates, A. Stolcke, K. Ries, D. Jurafsky, N. Coccaro, R. Martin, M. Meteer, and C.V. Ess-Dykema. 1998. Can prosody aid the automatic classification of dialog acts in conversational speech? *Lang. and Sp.*, 41(3-4):439–487.

E. Shriberg, A. Stolcke, D. Hakkani-Tür, and G. Tür. 2000. Prosody-based automatic segmentation of speech into sentences and topics. *Sp. Comm.*, 32(2):127–154.

M. Steedman. 2000. Information Structure and the Syntax-Phonology Interface. *LI*, 31(4):649-689.

H. Truckenbrodt. 2002. Upstep and embedded register levels. *Phonology*, 19:77–120.

E. Vallduví and M. Vilkuna. 1998. On Rheme and Kontrast. *Syntax and Semantics*, 29:79-108.