12

Annotating Texts for Language Documentation with Discourse Profiler's Metatagging System

Phil Quick

SIL International

This paper introduces a systematic and robust way to annotate (or 'tag') texts with discourse information. To date there has not been a method for annotating texts for language documentation with discourse-text information. This is the first paper to systematically describe the capabilities and the annotating methodology of the *Discourse Profiler's* metatagging system as a means of annotating endangered languages' texts in a *Toolbox* database. Since there is a division of labor between *Toolbox* and *Discourse Profiler*, the *Toolbox* database can be the basis for the archival tasks, whereas the *Discourse Profiler* software is a computer assisted discourse-text analytical tool that mines the *Toolbox* discourse-text annotated database in order to produce two primary capabilities: 1) to create a representative interactive compressed representation or 'map' of the structure and elements of a text, and 2) to quantify texts based on this special metatagging system with an array of sixteen different possible statistical outputs (including both referential distance and topic persistence statistics). Although the main focus of this paper is on the multipurpose annotation system, I will introduce the basics of the *Discourse Profiler* software in order to illustrate the range of analytical possibilities that this annotation system incorporates.

1. INTRODUCTION.[1] This paper introduces a systematic and robust way to annotate (or 'tag') texts with discourse information. There are two primary features of the *Discourse Profiler* software (available at www.discourseprofiler.com[2]). First, it produces an interactive, representative model of a text with a map-like abstraction. A key component of a text's map is the participant tracking which is represented with vertical lines, geometrical shapes for noun phrase types, and colors used to identify either the grammatical role or semantic role. Other vertical lines or 'spans' are used to trace the flow of information in a text parallel to the participant tracking map grid (i.e. span analyses). Second, it permits sixteen different possible statistical outputs, key amongst which are ten different topic continuity statistics that are automatically produced from the annotated text (including both referential distance and topic persistence statistics).

---

[1] I want to thank participants at the International Conference on Austronesian Endangered Language Documentation (June 2007, Taiwan) for their comments and feedback to an earlier version of this paper. I also thank the two anonymous reviewers for their input which has helped to improve this paper. My special thanks to Margaret Florey and D. Victoria Rau for some additional editorial help which goes a long ways toward streamlining my paper. I am responsible however for the presentation and any remaining problems, and I welcome further comments.

[2] *Discourse Profiler* is currently freely available at the website in its current beta 4.4 prerelease version. I welcome input and feedback on its development, including suggestions on the metatagging system. The current plan is to release version 1.0 as shareware.

Discourse Profiler has been developed to model and quantify texts based on a special metatagging system. This metatagging system was developed to work with any language, however the particular relevance focused on in this paper is on presenting this as an annotation system that can make a contribution to the documentation of endangered languages. To date there has not been a method for annotating texts for language documentation with discourse-text information.[3]

Although providing some basic information on discourse is clearly useful, the motivation for adding discourse-text information to the annotation of a text for language documentation should be made clear. Discourse-text information clearly adds to the overall richness of what we can learn about a text and its language. It is also the area of language that informs us about lower levels in the hierarchy of a language, especially in the domain of syntax (e.g. word order choices, voice selection criteria, transitivity identification, etc.), as well as intermediate levels such as paragraph structures, episode structures, and how propositional relations make a contribution to understanding the flow and texture of a text.

It is clear that a tagging system for a text should be usable for multiple purposes. A number of linguists have used Microsoft Excel as a means to do topic continuity analysis on various texts of individual languages.[4] The two serious drawbacks of these approaches are that 1) the 'tagging' (or annotating) method used has a single purpose and is for all practical purposes useful for only one time or one task, and 2) it is highly laborious. It is clearly more useful to have a metatagging system that has multiple uses because it is more efficient to annotate or tag something one time as opposed to multiple times. Even when the tagging system is laborious, it is more likely that a text will be tagged when the linguist knows there will be multiple uses to that tagging system.

A tagging system also needs to have constraints yet be as flexible as possible. For example, it is more flexible if it can handle multiple theoretical views of syntax. This includes being able to handle encoding all types of clauses (and even clause 'fragments'). Verbless clauses versus verbal clauses for example need to be differentiated yet to fit within the tagging constraints. The tagging system should not be too difficult to 'read', it should be fairly transparent, or at least easily mapped to allow for a simple interpretation process. For texts, it is also important to be able to track individual referents as well as 'plural' or 'mass' referents, e.g. 'they', 'the children', 'trees', etc. The tagging system should also not

---

[3] For example, it is not mentioned at all in Schultze-Berndt's excellent 2006 paper on annotating texts Schultze-Berndt (2006) does discuss 'discourse analysis', but this is with the meaning of 'conversation analysis'. *Discourse Profiler* was developed largely to deal with narratives and other lengthy monologues. This does not rule out that some conversation analysis could be done with some changes (possibly minor) in the annotation procedures. Rhetorical Structure Theory deals with some discourse-text information (see Taboada and Mann 2006). However as I understand it, it deals primarily with what I would call propositional relations. There is a developed annotation system for RST, however it appears generally more complex than what would typically be needed for an archival record.

[4] Cliff Olson (pers. comm.) has told me about a tool developed in SIL's Papua New Guinea Branch. I have heard of other people who have each developed their own custom approaches to doing topic continuity statistics on texts, all of which as far as I know were never used again.

be overloaded with information nor try to capture all information that is in a text. The tagging system should also be easily modified without a serious change in the parameters. It should also be a system that is easily implemented.

The metatagging system developed for the *Discourse Profiler* software fits all of these criteria. The strength of this metatagging system is that it already is multi-purposed for modeling texts and quantifying texts with the *Discourse Profiler* software package, yet can still be adapted, modified, or expanded for other uses. The tags contain a substantial amount of information as will be shown by the wide range of discourse analyses that can be used to model texts visually (with a large range of possibilities, incorporating ideas from Grimes 1975, Longacre 1983, 1996, Givón 1983, 1994, and Quick 1997 among others), and a variety of statistical approaches including a number of topic continuity statistical approaches (e.g. Dooley and Levinsohn 2001, Dryer 1994, and Givón 1983, 1994).

The following section introduces the *Discourse Profiler* software. Although the focus of this paper is on the metatagging system, it is important to understand what kinds of information can be analyzed using a multi-purposed systematic system for annotating a text. The introduction to *Discourse Profiler* demonstrates the robustness of this metatagging system. In Section 3, I introduce the two types of fields that are productive for annotating a text: information type fields and participant tracking fields. In Section 4, I discuss the main clause as the typical unit of description. The following section then briefly examines three features in Toolbox that especially are of help in annotating a text or working with the finished annotated text. In Section 6, the use of *Discourse Profiler* is illustrated through its application with various grammatical categories drawn from the endangered Pendau language (Sulawesi, Indonesia). The conclusion highlights the benefits of *Discourse Profiler* for the documentation of endangered languages and discusses a number of features to be developed in the future.

2. *DISCOURSE PROFILER*: A COMPUTER TOOL FOR MODELING AND QUANTIFYING NATURAL LANGUAGE TEXTS. Linguistic work on language texts has until now lacked a serious software tool that integrates basic linguistic theory and discourse theory. The *Discourse Profiler* software package provides linguists with a tool to analyze the entire context of specific discourse features.[5] These features are summarized here. Practical applications and benefits of using the *Discourse Profiler* program have already been implemented in Quick (2002, 2003, 2005, 2007) for the Pendau language (Sulawesi, Indonesia). Examples of use of the program with Pendau are included in this section.

2.1 A ROADMAP FOR TEXTS. A visual model of a text is analogous to a roadmap. Different size cities are represented by an iconic change in size. In a text, different noun phrases can be identified by a different shape, such as using circles and squares to contrast nominative case and accusative case. Colors in roadmaps are often used to contrast geographical features such as blue for water. Colors in *Discourse Profiler* are used to represent semantic or grammatical details such as red for semantic agent, and black for semantic patient. In addition to the basic participant tracking that is displayed in map-like format, syntactic and/or discourse information such as same subject/different subject, event/non-

[5] Earlier beta versions were called *Multilinear Discourse Analysis (MDA)*. The first demonstration version was demonstrated in 1996 (see Quick 1996).

event, and word orders can be traced parallel to the basic text's map display. This allows for analysis of a range of possibilities and the interactive capabilities adds the further help of trying 'what if' easily and rapidly for shifting to various hypotheses and to eliminate or elucidate patterns.

The visual model or map of a text allows the linguist to view fifty clauses and thirty participants in a single view, and therefore clusters of information can be easily compared to the participant tracking. These comparisons and flexible changing of settings (e.g. change the color of subjects from blue to black) allows the linguist to identify possible patterns that are often difficult or impossible to determine when analyzing the actual text.

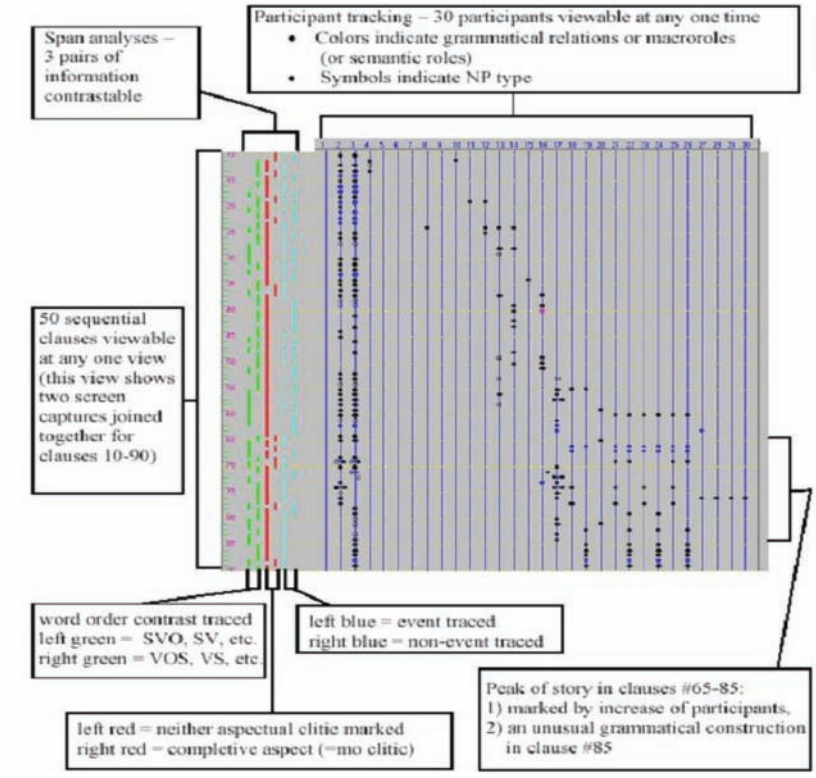


FIGURE 1. Abstract profile of the peak of a Pendau folktale (clauses 10-90) in the *Discourse Profiler* program

With visual modeling of Pendau data, I was able, for example, to compare the span analysis of word order (usually SV or SVO), the occurrence of completive aspect, and the occurrence of certain temporal relators (e.g. *ila uo* 'after that'). When a clause has a participant change, I could confirm where the boundaries of paragraphs in Pendau occurred most frequently (Quick 2003, 2007).[6]

Figure 1 shows a screen view of clauses 10-90 (top to bottom) of a short recorded text as an abstract visual model or profile. It provides a view of the participant tracking of participants 1-30 (from left to right). Participants are the various geometrical shapes (dots, squares, circles, etc.). This view is a 'map' of a text and allows the user to interact with the syntax and discourse in many different ways. The main part of the 'map' provides a means for participant tracking. The user assigns different symbols to a NP type, for example a circle might represent a pronoun, and a square might represent a basic NP. Colors can be used to track grammatical relations, e.g. red might indicate a grammatical object, and green might indicate a grammatical subject, or semantic roles (or macroroles) can be contrasted likewise. On the far left are pairs of vertical lines used for span analyses (Grimes 1975). These allow the discourse analyst to track or trace discourse and/or syntactic level information that parallels the participant tracking. For example, event and non-event can be contrasted parallel to different word orders, e.g. contrasting SV/SVO with VOS/VS.

Another well-known literary/discourse feature that occurs in the story profiled in Figure 1 is the gathering of a lot of participants in the peak (see Longacre 1983, 1996). The two main participants are easily identified by participant tracking lines number 2 and 3. Other participants appear fairly randomly until we get near the bottom of this screen view (which is also near the end of the folktale). The zone of turbulence that occurs in a discourse peak can actually be seen visually here in the preceding clauses as the number of participants increases for a number of clauses between 65 and 85. The peak of this narrative has an uncommon grammatical construction (in clause 85) which has not been documented anywhere else in my corpus, but when checked in elicitation it was not considered to be at all unusual.[7]

2.2 METATAGGING. The annotation system includes a metatagging system that allows an integrated approach to text analysis and/or text annotation. The entry for data in Discourse Profiler is expedited by utilizing the *Toolbox* software that many linguists are already familiar with. Although the term 'metatagging' can generally be thought of as synonymous to 'annotations' it really is distinct from the typical annotations that, for example, Schultze-Berndt (2006) describes. This is because the two field types are actually mini databases (note especially the participant tracking fields with its five 'tiers'). However, since

[6] This interaction and visual inspection of the map was only part of the process for identifying paragraph boundaries. There was also some statistical analysis performed on the occurrence and context of the completive aspect *=mo*. Paragraph boundaries may also include changes in location, time, and a change in the setting as well as such things as the beginning of direct speech.

[7] Longacre (1983, 1996; also see Edmondson and Burquest 1992:84-90) describes a number of linguistic signals that indicate a narrative's peak. This story has at least two of these signals: an increase of participants and an uncommon grammatical construction.

it does fit in generally with the basic notion of annotating at least from the user's perspective, I adopt the term annotation when the focus is on working in *Toolbox*, and use the term *metatagging* when I am focusing on using the *Discourse Profiler* software.

Toolbox (and its predecessor *Shoebox*) has become a significant software tool for developing and maintaining lexicons and texts, and for outputting dictionaries. The interlinearizing feature has helped linguists become more productive with its semi-automatic feature of building up a lexicon through the interlinearization process. Although the methodology presented here can be used independent of *Toolbox*, I will assume for ease of discussion that the majority of linguists who will adopt this annotation methodology will at least be using *Toolbox*, and likely will want to use *Discourse Profiler*.

Figure 2 shows a flow diagram of the relationship between *Toolbox* and *Discourse Profiler* and how these annotations are used in each. Data entry of the annotation fields is carried out only in *Toolbox*. The separation of the text database from *Discourse Profiler* is an important one as it allows archiving of plain texts without any interference from *Discourse Profiler*.[8] The flow diagram shows that there are two types of annotation fields: 1) information type fields, and 2) participant tracking fields. These will be further explained

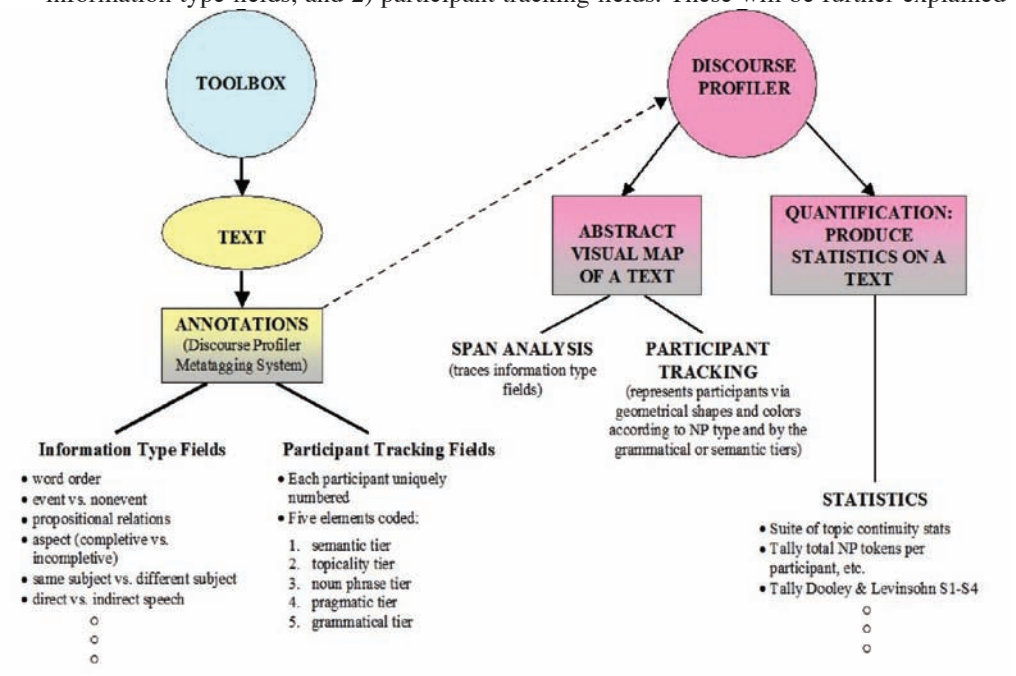


FIGURE 2. Flow Diagram of How the Metatagging System is Used in Toolbox and Discourse Profiler

[8] This also allows other linguistic analysis to be carried out in *Toolbox*, since certain grammatical and discourse information will then be available.

in later sections, but it is important to note that the data entered as annotations here are multi-purposed in order to serve multiple analytical possibilities. After these annotations are entered (also referred to here as 'metatagging') along with other language annotations the linguist may be making, the database can be: 1) archived and 2) further analyzed at any time with *Discourse Profiler* or some future software.

The only link between *Discourse Profiler* and *Toolbox* is when *Discourse Profiler* mines these multi-purposed annotated data. This mining only extracts information and does not adjust the *Toolbox* database in any way.

One of the benefits of the *Discourse Profiler* program is that it allows the user to make rapid shifts and/or refine one's hypothesis or analysis, and to analyze a higher number of texts more efficiently. This flexibility is made possible by the metatagging system and their use within the span analysis settings feature of *Discourse Profiler*. For binary information such as the contrast of event and nonevent, these are simply 'listed' in two separate lists so that it is easy to trace the occurrence of each one on separate span 'lines' in the map of the text. Information that has more than two items is grouped into two different lists.

For example, in Pendau I list the two contrastive word orders of SV/SVO and VOS/VS (and many other variations which may include obliques and clauses with zero anaphora) into two groups (or lists). These two separate lists of the word orders allow one to work on hypotheses to determine why there is a variation in this word order difference. These groupings are easily changed in these lists so that the span analysis can be compared with other spans and with the text's participants. Likewise, the settings for participants allow one to reassign the colors and/or shapes easily and quickly. This allows for a great degree of flexibility in increasing the possible analyses. All of these settings can be saved for different analyses of the same text, or using one or more of these settings for as many other texts that will be analyzed. By using the *Discourse Profiler's* settings there will be nothing in the original *Toolbox* database that will be changed.

2.3 STATISTICS. A range of statistical options often used to analyze texts manually is now automated. The time saved can thus be put into analyzing a larger number of texts. The quantification of texts ranges from various topic continuity statistics (e.g. the Givón 1994 approach is different from the Dryer 1994 approach) to basic statistics on the number of noun phrases for each participant tracked.

Statistical analysis was undertaken for the topic continuity of the core arguments in two different transitive constructions in Pendau. This analysis provides evidence that both of these constructions are equally transitive. Active and inverse verbal clause constructions are nearly equal in frequency of occurrence. Discourse topic continuity studies show that the A argument in inverse voice clause constructions is highly topical in this language, and a comparison of the A and P arguments in both active voice and inverse voice clause constructions have a similar profile as expected for transitive clauses. Table 1 illustrates this with one of the texts (Mtext3) from which the statistics were generated with *Discourse Profiler*.[9]

[9] The generalizations were gleaned from four texts (Quick 2003, 2005, 2007)

	Inverse Voice	*Inverse Voice*	*Active Voice*	*Active Voice*
RD	P	A	P	A
1-3	70 (70.71%)	65 (82.28%)	24 (53.34%)	79 (89.77%)
>3	29 (29.29%)	14 (17.72%)	21 (46.76%)	9 (10.23%)
Total	99 (100%)	79 (100%)	45 (100%)	88 (100%)

TABLE 1. Referential distance values in Pendau—Mtext3

Topic continuity statistics also suggest that in Pendau previous discourse information is important in the speaker's choice between active voice and inverse voice. A matrix such as shown in Table 2 is produced with the raw data of each text analyzed. The data can then be copied into a text editor for better formatting as illustrated with Table 2. In the case of Table 2, the data from the four texts (Mtexts 1-4) were added together in order to create a final version. Table 2 illustrates the version of topic continuity statistics developed by Dryer (1994).

	Inverse Voice	Active Voice	Total
RD of A lower	60 (38%)	100 (62%)	160 (100%)
RD of A and P same	83 (72%)	33 (28%)	116 (100%)
RD of P lower	100 (60%)	67 (40%)	167 (100%)

TABLE 2. Relative referential distance of As and Ps (Dryer horizontal analysis)
(All texts combined—Mtexts 1-4)

2.3.1 GRAPHING. The statistics produced for each of the four Pendau texts can also be plugged into a graphics program (Figure 3).[10] As typically produced for topic continuity studies, a scatter plot graph is used. For the Pendau data, if the referential distance of the Undergoer is less than the distance for the Actor within the same clause (P<A) or the referential distance for the Actor and Undergoer of the same clause is the same (P=A), then the inverse voice verbal construction will more often be chosen. But if the referential distance of the Actor is greater than the Undergoer in the same clause (A<P), then the active voice

[10] All of the statistics produced are raw data. If the linguist wants to make graphs, then s/he will need to use a graphing program. This usually just involves plugging in the raw numbers, and then producing the graph. All of the matrices produced as raw statistics in *Discourse Profiler* can be copied and entered into a table for better formatting. I have been able to format and graph all of the statistics on Pendau using MS Word.

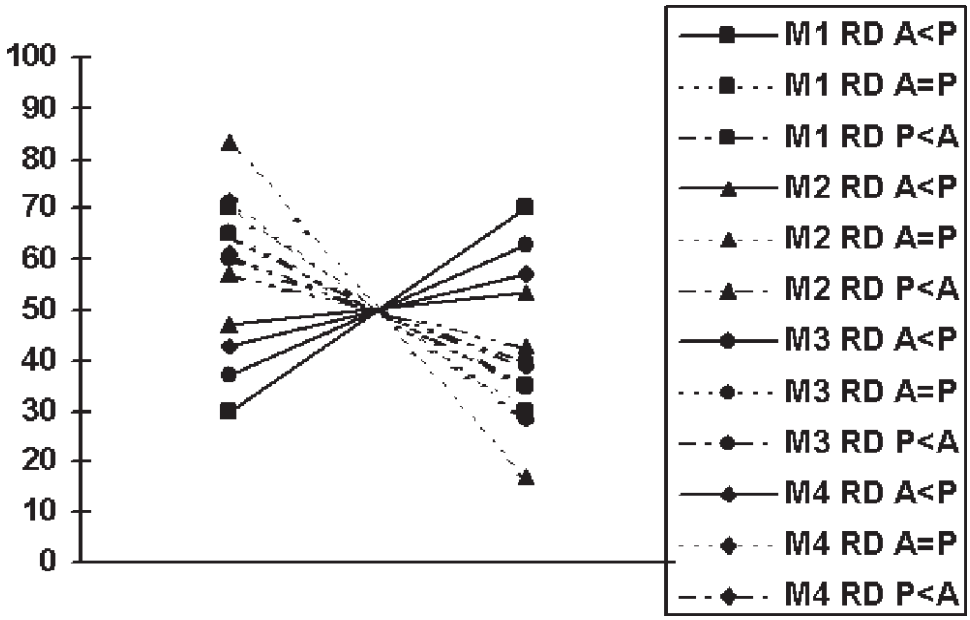


FIGURE 3. Frequency of A and P arguments in active and inverse voice constructions according to whether the A is equal to P in referential distance (RD A=P), the A is less than P in referential distance (RD A<P), or the P is less than A in referential distance (RD P<A); the data are from four Pendau texts, labeled as M1-M4

verbal construction will more often be chosen (see Figure 3). Figure 3 shows there are four areas of significant clustering that occur in the scatter plot graph (roughly in each of the quadrants; M1-M4 refer to four analyzed texts, in which the M reflects the previous beta versions' name for *Discourse Profiler*, which was *Multilinear Discourse Analysis*).

Another basic but potentially highly useful statistical approach is to tally the number of participants according to noun phrase types in various ways. Figure 4 shows one approach to tallying referents of a text with a bar graph. This shows how many occurrences of each participant in a text occur as a different type of noun phrase. For example, in the text quantified in this bar graph it is easy to ascertain that the most common participants are participants one and two (the numbering of the participants is from left to right within each noun phrase type). It is also noted that participant one (red bar on far left of each NP type) appears most frequently as a genitive case pronoun, followed by zero anaphora, and then as a typical noun phrase in absolute case and then in the genitive case. Participant two (brown bar, second from left in each NP type) appears very frequently as a typical noun phrase in the absolute case and very high frequency in zero anaphora. Its other occurrences are quite low.

Two other ways to make similar tallies include how each participant occurs as subject, object, oblique and which type of noun phrase these occur as. Different tallies of a text can be done to compare different cases, participants with relative clauses, where demonstra-

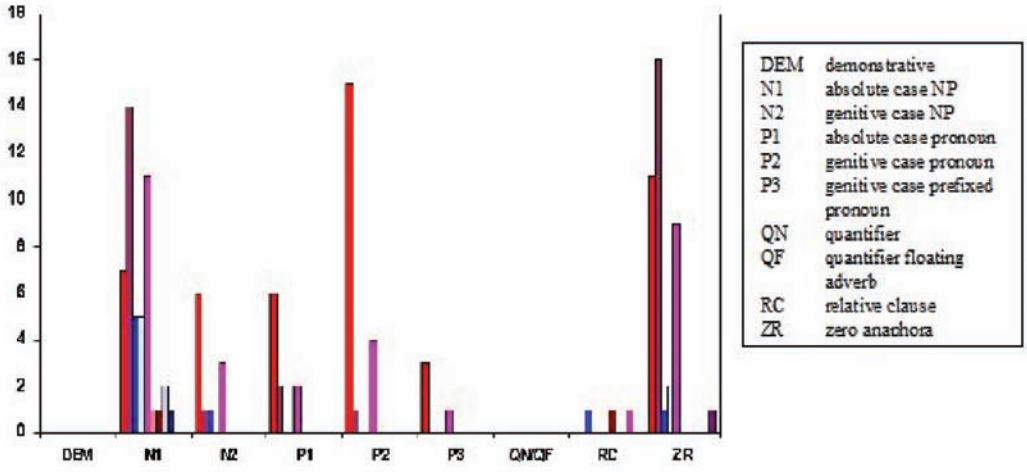


FIGURE 4. NP profile for participants 1-11 in one Pendau text (each color represents a different participant, e.g. red is participant one for each NP type it occurs in)

tives occur, etc. For example, in Pendau typical nouns and pronouns are coded as N1 and P1 respectively for absolute case, and as N2 and P2 for genitive case. Rather than tally these occurrences for each referent they could be tallied for total occurrences in absolute case versus genitive case. This would show in a matrix how many times each participant appears as a pronoun in absolute case, as a pronoun in genitive case, as a typical noun phrase in the absolute case, and as a typical noun phrase in genitive case.

2.3.2 QUANTIFICATION TOOLS. The list below summarizes the quantification tools currently available in *Discourse Profiler*. Altogether there are eleven different possible matrices and five sets of matrices of quantified data that can be produced (the sets of matrices are dependent on how many annotations a particular tier has).

Create tally of total NP tokens per participant. This produces a matrix showing how many times each participant occurs for every NP type that has been annotated (e.g. basic NP, pronoun, zero anaphora, etc.)

Create tally of total NP tokens according to the semantic tier tags. This produces a matrix showing how many times each NP participant occurs as a specific category annotated for the semantic tier (e.g. how many times an NP is agent/actor, single argument, patient/undergoer, second object, etc.).

Create tally of total NP tokens according to the grammatical tier tags. This produces a matrix showing how many times each NP participant occurs as a specific category annotated for the grammatical tier (e.g. how many times a NP is subject, object, left-dislocation, second object, oblique type, etc.).

Create basic Dooley and Levinsohn S1-S4 statistics for each NP type. This creates a matrix that generalizes the statistics for all the participants. The matrix shows how many occurrences for each of the S1-S4 categories there are for each NP type.

Create Dooley and Levinsohn S1-S4 statistics for each participant according to all of the possible NP types that have been annotated. This produces a different matrix for each participant. Each participant's matrix shows how many occurrences there are for each of the S1-S4 criteria according to each possible NP type. So if a text has eleven participants, then there will be eleven participant matrices.

Create Dooley and Levinsohn S1-S4 statistics for each participant regardless of the NP type. This produces one matrix listing all participants and how many occurrences there are for each participant and for each of the S1-S4 criteria.

Perform basic topic persistence (TP) and referential distance (RD) statistics based on semantic tier tags. This produces one set of matrices for the TP and one set of matrices for the RD for each annotation used in the semantic tier (e.g. actor, undergoer, oblique category, etc.). One set of TP matrices and one set of RD matrices is produced for each semantic annotation. So the number of matrices depends on how many categories have been annotated.

Perform basic topic persistence (TP) and referential distance (RD) statistics based on grammatical tier tags. This produces one set of matrices for the TP and one set of matrices for the RD for each annotation used in the grammatical tier (e.g. subject, object, oblique, etc.). One set of TP matrices and one set of RD matrices is produced for each grammatical annotation. So the number of matrices depends on how many categories have been annotated.

Perform advanced topic persistence and referential distance statistics using the Givón method that compares two probable or possible transitive constructions simultaneously. This produces one TP matrix and one RD matrix.

Perform advanced topic persistence and referential distance statistics using the Dryer method that compares two probable or possible transitive constructions simultaneously. This produces one vertical TP matrix, one horizontal TP matrix, one vertical RD matrix, and one horizontal RD matrix. This method compares the frequency of the actor and undergoer within the same clause according to whether they are the same or lower in frequency.

An additional benefit for the *Discourse Profiler* tags in the *Toolbox* database can be to search for specific examples of text data with a specific code or coding combination. These can also be used for various grammatical analyses and for finding examples to use in research papers or for other research goals (see Pastika 1999, 2006 for examples of an application of this metatagging methodology used in his Balinese texts).

3. FIELD TYPES. *Discourse Profiler* includes two field types: information type fields and participant tracking fields. These are described in the following sections.

3.1 INFORMATION TYPE FIELDS. It is helpful to record some kinds of linguistic information which can be traced parallel to participant tracking. The idea for the information field originates from Grimes' (1975) description on how to do a 'span analysis' of various discourse information. A span analysis is similar to participant tracking, but is a method of tracing the information flow via spans of information as they appear in clauses. Comparing different spans that appear together, and as they correlate with the participants that are being tracked offers a methodology of identifying patterns where information clusters together. This may help to determine for example where paragraph boundaries occur, or a number of other possibilities.

The information annotated in these fields is typically typological or discourse oriented. Each category of information type must have its own respective field, however the user may have as many different fields as information types that s/he would like to analyze. Typical information types that would be recorded in their own fields are listed below:

- Aspect (e.g. completive versus incompletive; other categories of TAM)
- Clause type (e.g. declarative, interrogative)
- Conversation analysis (this is somewhat limited as the *Discourse Profiler* is designed primarily for use with narratives)[11]
- Dependent clauses (e.g. adverbial clauses, temporal clauses, peripheral elements)
- Direct speech versus nondirect speech (e.g. quoted material)
- Discourse category (e.g. following Grimes (1975), event, nonevent, setting, narrator evaluation)
- Phonetic/phonological features (e.g. loud, soft, aspirated, vowel harmony)
- Propositional relations (connectors; e.g. but, therefore, since)
- Repetition (e.g. tail-head linkage, resumptive repetition, iconic repetition)
- Same subject versus different subject (e.g. Papuan, South American language features)
- Transitivity (e.g. Thompson and Hopper's (1980) categories)
- Unit types (e.g. paragraphs, conversational turns, intonation units (IU) that span more than one clause unit can be marked as in IU1a, IU1b, IU1c, IU2a, IU2b)

[11] Turn taking and identification of speakers of a conversation can be 'tracked' through the information type fields (i.e. span analysis) fields. For example, one could trace up to eight speakers by using four fields such as \sp1_2, \sp3_4, \sp5_6, and \sp7_8. The identification of speakers is then given in the field for a clause, as in '\sp1_2 speaker2' for a second speaker (or by using the actual name). In the settings within *Discourse Profiler* then for each pair of speakers, only one of the pair of speakers is entered in its own 'list' (i.e. only one item is 'listed').

- Unusual grammatical features (versus typical)
- Verb Spectrum Profile (e.g. Longacre's approach to discourse analysis (1989))
- Verb types (e.g. specific verb classes in languages which systematically differentiate classes following morphological criteria such as stem formers in Pendau)
- Word Order (e.g. SVO, VOS, SV, VS)

The fields may be commonly abbreviated \it1, \it2, \it3. Alternatively, for information that is widely known to be useful in many languages such as word order, fields can be abbreviated mnemonically: e.g. \wo. The information normally recorded in these fields is typically contrastive binary information such as 'same subject' versus 'different subject' or information that is normally grouped together into potentially similar or contrasting groups as in various word orders. Typical abbreviations, words, or abbreviated words are entered in these fields. In fact, this information is of a type that is often already entered by analysts. Additionally, specific information found only in an individual language can also be entered in these fields. Multiple fields are used, but the information type for a particular category is restricted to its own respective field.

In summary, there is no closed number to the information type of fields that can be used and the linguist may freely use as many fields as s/he wants to for typological or discourse information as one would normally use within Toolbox. This will most likely be dictated by practical concerns and areas of interest typically worked on in linguistics. This is therefore the more flexible of the two field types. The only constraint is that each category of information must be restricted to its own field. This constraint allows the program *Discourse Profiler* to trace each category of information parallel to the participant tracking in the visual display of a text. This trace or span analysis is useful to locate patterns that are otherwise difficult or impossible to analyze through conventional methods.

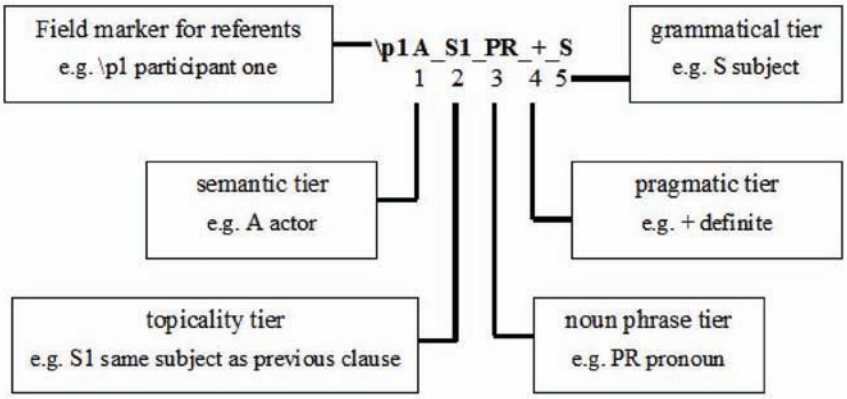


FIGURE 5. The five tier positions used in the participant tracking fields (exact sequence provided by the linguist)

3.2 PARTICIPANT TRACKING FIELDS AND THEIR FIVE ELEMENTS. In *Discourse Profiler,* each referent must be uniquely identified with its own field. This is done by using the letter 'p' followed by a number: for example, 'p21' for participant number 21, and following SIL's standard format system appears with the backslash marker then as: \p21

There are five elements that need to be coded into one field for each participant (i.e. as requirements of the *Discourse Profiler* software), see Figure 5. These five elements will be referred to as 'tiers'. These tiers contain autonomous information typically and reflect categories commonly used by the majority of linguists. The information in each of these tiers is abbreviated typically according to the user's own needs and allows latitude for differing theoretical approaches. The user needs to remember that these tiers are not fields for annotation purposes, and so cannot be left blank. Information type fields can be left blank when there is no relevant information to annotate for a particular clause.

Typically the information encoded for each of the five elements will be an abbreviation (open to the linguist) rather than a word. This is largely because the information within the field needs to be kept down to a practical length (typically one to five characters is enough for each tier). This is primarily for readability reasons, and secondarily for computational reasons. The choice of the abbreviation used for each tier is left up to the linguist (except for use of the topicality tier), although as with many other descriptive methods this needs to be consistent. The important point is to keep the abbreviations for each of these five areas in the same sequence (i.e. 12345, or 54321, etc). Of the five tiers, only the NP type tier is open as to the set of NP types identified. The other four tiers are fairly restricted to what is available to identifying the particular participant as determined by the syntax of each clause. Compare Figure 5 with the following descriptions of each tier.

3.2.1 SEMANTIC TIER. For verbal clauses this tier is reserved for the basic macroroles actor and undergoer, or what are often referred to as A and P arguments. Single argument clauses can be further delineated as is commonly practiced in linguistics with the capital S, as the S, A, and P are often the means used to identify differing grammatical systems. This tier is not really meant to be used for semantic roles (e.g. experiencer, instrument, etc.), although with some modification it can work. The reason for this constraint largely has to do with how the calculations of advanced topic continuity statistics are carried out in the *Discourse Profiler* software. For intransitive clauses it may be useful to indicate the difference between undergoer and actor single arguments, as in Su or Sa respectively. For nonverbal clauses this is the tier to use to identify the nominal argument as simply the first or second argument.

3.2.2 TOPICALITY TIER. This tier is for a special category of topic continuity in which Dooley and Levinsohn (2001) discuss the importance of whether an NP is activated or not within a particular context. This can be done for the subject (S1-S4) or for the non-subject (N1-N4). The range for subject is as follows (Dooley and Levinsohn 2001: 130):

- **S1** the subject is the same as in the previous clause or sentence
- **S2** the subject was the addressee of a speech reported in the previous sentence (in a closed conversation)

- **S3** the subject was involved in the previous sentence in a non-subject role other than in a closed conversation
- **S4** other changes of subject than those covered by S2 and S3

This is a technique that Dooley and Levinsohn (2001) have developed that comes partly out of treatments on topic continuity (e.g. Chafe 1987, Givón 1983, 1990) and partly from the topic-comment literature (i.e. focus, topic and sentence articulations, e.g. Andrews 1985, Chafe 1976, Givón 1990). The purpose of this technique is to identify the amount of coding material used in each category (especially for S1-S4), and then to determine what coding material is used and the 'motivations for deviances from default encoding' (Dooley and Levinsohn 2001: 134). They state for example that (2001: 134),

> …common motivations for increased encoding include the presence of a discontinuity and the highlighting of information, while decreased encoding is typically used to identify a VIP.

If this tier category is of no interest then this can be considered to be an optional tier, but due to the constraints currently in *Discourse Profiler* this must be filled in by a dummy character such as with an asterisk. At this stage in the development of the software, other information could be used by the linguist (i.e. annotated here), but there is little point of doing that as there is nothing that *Discourse Profiler* would be able to process.

3.2.3 REFERENT'S NOUN PHRASE TYPE IDENTIFICATION TIER. This tier answers the question: what form does the noun phrase take? If it's a simple noun phrase, then a common abbreviation such as NP may be used. If it's a basic pronoun, then some other simple abbreviation such as PR may be used. If a case system is used, then another abbreviation can be devised to contrast the noun phrases as such. If a noun phrase is omitted then it is marked as such, for example, if it is due to simple zero anaphora, then ZR may be used to abbreviate it. The choice of abbreviations here can be a rather small set or rather complex depending on the needs of the linguist.

3.2.4 PRAGMATIC TIER. This tier answers the question: is the NP definite or indefinite? Alternatively this tier can be used for givenness, specificity, old/new information, etc., however only one category can be used. Definiteness should be the primary consideration here. If the linguist also wants to annotate more than one of these categories, then s/he has the option to use the information type fields and use the tracing feature in *Discourse Profiler*. Technically this is also an optional field, so if one does not want to identify any of these parameters, a dummy character can be used for this tier.

3.2.5. GRAMMATICAL TIER. This tier answers the question: what is the grammatical relation (or pivot/non-pivot, etc.) for this participant? This tier typically distinguishes the grammatical subject, object, indirect object, second object, etc. For nonverbal clauses this is the tier reserved for identifying the grammatical function of the noun phrase, typically either as the subject or as the predicated noun phrase. For some languages it may also be useful to annotate the intransitive subject differently from the transitive subject.

These five tiers are coded for each participant of each clause–normally only 2 to 3 referents (and or props) will be coded for each clause. Example (1) provides a typical description of a record with a brief description of what occurs in each field that would typically be used by a linguist for a text. The fields that immediately follow the vernacular text line (\txt) are typically used in interlinearizing a text (e.g. \mr, \ge, \ps). Glosses that occur in the part of speech line (\ps) may sometimes have the same abbreviation used in the NP tier, however this is usually minimal redundancy when and if this occurs. Since the interlinearized portion is usually produced semi-automatically, the requirement to also identify the NP type in the NP tier for each participant is one of little additional time. In the participant tracking fields note that the underline character is used to separate the five different tiers. It is helpful to separate each of the tiers for readability reasons and for computer processing reasons (it will be required once *Discourse Profiler* is released as version 1).

(1)

\ref Text 001	Required record marker
\txt Vernacular text goes here	This is not necessary for Discourse Profiler and is only necessary if the user is working with a Toolbox database
\mr morpheme break line used in Toolbox for example	Used for interlinearizing; optional—not necessary for Discourse Profiler
\ge English gloss	Used for interlinearizing; optional—not necessary for Discourse Profiler
\ps Part of speech	Used for interlinearizing; optional—not necessary for Discourse Profiler
\fte free translation	Optional—not necessary for Discourse Profiler
\wo SVO	Word order of clause—optional in Discourse Profiler but usually for Span Analyses
\it1 Event	Discourse information type 1—contrast for example event and nonevent in the Span Analyses—optional in Discourse Profiler
\it2 SS	Discourse information type 2—contrast for example Same Subject versus Different Subject in the Span Analyses—Optional in Discourse Profiler

\p2 1_2_3_4_5	Necessary for each participant in the clause—numbers represent the five tiers that are used for the abbreviations, e.g. participant 2
\p5 1_2_3_4_5	Necessary for each participant in the clause—numbers represent the five tiers that are used for the abbreviations, e.g. participant 5

Example (2) illustrates a record used for a clause from a Pendau folktale.[12]

(2) \rf fktale01.txt 002b
\pen Ila uo jimo asi mene' negutu sanu binaung.
\mr ila uo jimo asi mene' N-pe-gutu sanu binaung
\ge ABL yonder 3PL/GE just go_up RE-SF/DY-make umm lean-to

\fte After that they just went up to make umm a lean-to.
\wo SVO
\it1 Ila uo
\it3 event
\p2 A_*_P1_+_S
\p3 A_*_P1_+_S
\p4 A_*_P1_+_S
\p7 P_*_N1_-_O
\dt 07/Apr/2000

In this example the word order is SVO. Information type field one occurs after the word order information type field and is used in this text to identify the discourse connectors (or relators that are used to identify the particular propositional relation between preceding and subsequent information). Next, information type number three identifies the discourse information as an event clause. Information type two is omitted as it is not relevant for this clause. The information type fields are followed by four participants. Participants \p2, \p3, and \p4 are three men in this story who have been previously distinguished as distinct referents. Since they are referred to by a plural pronoun *jimo* 'they', they are all identified identically with the same five tiers, but identified as distinct by giving them separate numbered fields (see §6.1 for different ways to code plural referents). They are all 'actors', and marked with a plus to indicate they are definite. The asterisk indicates that the topicality field is marked with a dummy symbol, i.e. an asterisk in this case. The P1 indicates this is the pronoun used from pronoun set one (or the absolute case). The S indicates that the

[12] Abbreviations used in the Pendau interlinear glossing are as follows: 1SG first singular, 3PL third plural, 3SG third singular, AB absolute case, ABL ablative, COMP completive, DY dynamic verb class, GE genitive case, IV inverse voice, LOC locative, RE realis, RM relative marker, SF stem former, ST stative verb.

referent is the grammatical subject. Referent seven is from noun set one (absolute case), and is indefinite as it is introduced here for the first time in the story. It is an undergoer that functions as the grammatical object.

4. TYPICAL UNIT OF DESCRIPTION. An important point which must be made is that the typical unit of description is the main clause or what is typically understand to be a 'simple sentence'. This follows standard practice in descriptive linguistics, and coincides with the choice made in Rhetorical Structure Theory to use the clause as the "elementary discourse unit" (Carlson and Marcu 2001, also see Taboada and Mann 2006). Another important reason for making the basic clause the standard unit is that this works best for participant tracking and follows the methodology practiced for topic continuity analysis (e.g. Givón 1983, 1994). Each database record will typically be a clause, however this does not mean that these units cannot be relative clauses or other subordinate clauses. This is partly a matter of descriptive choice, and partly a matter of practical concerns that are not always easily resolvable. Gildea (1994: 208-211) discusses some of the guidelines needed to determine what is a clause for topic continuity analyses. Difficult decisions often need to be made that often parallel the difficulty of a technical theoretical analysis. On the other hand, because the annotation procedure does not require a rigorous theoretical analysis some room can be made for more practical judgments.

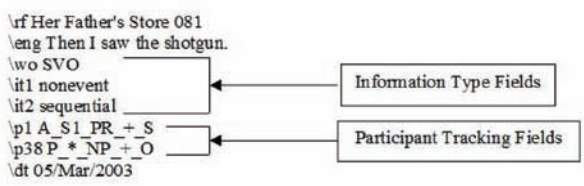
Clausal fragments or interjections often occur in texts and may be incomplete or independent from a syntactic clause. These can still be entered in as if they are a clause unit. If there are participants, these can be tracked through the annotation system. If the interjection is simply a statement by a participant such as 'yes', then no participant tracking fields will need to be used. One can still note in the information type fields who the speaker is, if one wants to trace a limited number of speakers. Some adverbial clauses, such as a temporal clause that is dependent on the matrix clause are usually left with the main clause. If one needs to trace adverbial information, then this can also be annotated in a dedicated information type field.

Although the constraint of using syntactic clauses is to be used as the general guideline, the annotation system is flexible enough that it is not mandatory. For example, it is possible to change the units to intonation units or to include sentences such as an if-then sentence that would have two syntactic clauses. This is acceptable to use occasionally, however it is better to indicate the intonation units in the information type fields (or as generally practiced, this can also be indicated in some part of the interlinearized text; see Schultze-Berndt 2006, Himmelmann 2006). If the linguist wants to trace intonation units, this can then be done in the interactive visual map in *Discourse Profiler*. As for sentences with a propositional relation such as an if-then or cause-effect propositional relation, it is better overall for the purposes of the statistical algorithms in *Discourse Profiler* to break these into two clauses, and then trace the semantic components of each clause in an information type field as to its propositional relations. Another related reason for breaking up complex clauses into separate syntactic units is so that the participants can be tracked sequentially or chronologically. If a complex clause has simultaneous actions occurring between different participants, this could perhaps be an exception, and then the participants would all be coded in the same unit. However even an exception like this could potentially skew some of the topic continuity statistics.

The tagging approach is simple as it entails adding only a few extra fields of information to each record in a *Toolbox* database, however it allows for a range from simple to complex ways of coding information concisely. The metatagging system has two types of fields that are important to the linguist: 1) information types (e.g. contrasting event/nonevent, word order, and any number of other grammatical features or discourse information), and 2) participant tracking information for all major referents (this requires 5 elements within each participant/referent field: identification of an NP type, semantic role, grammatical relation, definiteness or specificity, and topicality status).

The example in (3) shows one record from an English fiction story (Quick 1977) with the two types of fields labeled.

(3)



The first two fields identify the record, and contain one clause/sentence from this text. Interlinearized text and other annotated fields could occur here as well (see Schultze-Berndt 2006). The information type fields each abbreviate or summarize the particular grammatical or discourse type of information that the linguist wants to document. In this example these are word order (\wo), event versus non-event information (\it1), and propositional relations (\it2). The word order set includes the full range of word orders possible in a language. Although I use a binary set for tracing whether the discourse information in information type one in the example text above is either an event or a non-event (similar to foreground and background), this can be given a more detailed set if one wants to follow for example Hopper-Thompson's set of ten transitivity features (1980) or follow a different approach such as the set from Grimes (1975; Dooley and Levinsohn 2001: 81-83) which includes: event, participant orientation, setting, background (i.e. explanation), evaluation, performative, and collateral. The propositional relations in Pendau for information type two include: cause-effect, concession-contraexpectation, condition-consequence, simultaneous, overlap, alternation, and sequential.

The following fields demonstrate the participant tracking fields. Each referent or participant receives its own unique number, marked in this example as \p1 and \p38. In this sentence participant one is the narrator and is marked with codes which capture these five elements of information: actor/agent (A), same subject as previous clause (S1), pronoun (PR), definite (+) and grammatical subject (S). The shotgun referent's code can be interpreted as: patient/undergoer (P), unmarked for topicality (*; or not applicable), a typical noun phrase (NP), definite (+) and the grammatical object (O).

Although this metatagging system was developed for the *Discourse Profiler* software, it is not necessary to use this software in order to take advantage of it for annotating or archiving texts. The metatagging system also takes advantage of the capabilities of *Toolbox*, however it is not necessary to use *Toolbox*. The choice remaining would be to use a word processor (or basic text editor) that can save the data in plain text format. This extraordi-

nary choice would still necessitate using the field markers in the database format developed by SIL, and as used in *Toolbox*. Although this is not the preferable way to work, there may be some circumstances that preclude using *Toolbox*. This approach then still allows the linguist or language worker to annotate texts for documenting a language for archival purposes, and allow the possibility of further analysis of a tagged text using *Discourse Profiler*. There may well be another software available in the future or one that could be adapted to using the kind of database structure used in *Toolbox*, and this as well would likely be for exceptional reasons.

5. TOOLBOX AND DISCOURSE PROFILER. There are three features in Toolbox that especially are of help in annotating a text or working with the finished annotated text. These features are the semantic range feature, the browse mode feature, and the filtering feature.

5.1 SEMANTIC RANGE. This feature is helpful for staying consistent for delimiting the range of abbreviations or words used in a particular field while one is entering the annotation. It works similarly to a spell checker, and not only allows the words or abbreviations already kept in a special list for that particular field, but allows new additions easily as needed. This feature can be particularly useful for the participant tracking fields, as there may easily be twenty or more possibilities coded for a participant. Once an initial list of possibilities is listed for a particular participant, the semantic range feature provides a running list that can be used as a menu to choose from. This means there is less typing to perform.

5.2 BROWSE. This mode is helpful for doing some basic discourse analysis, in addition to typical syntactic analysis often done for descriptive purposes (also useful when looking for examples to use in a paper). For example, in the browse mode the user can view multiple records with selected fields displayed in columns. This feature allows one to do some basic span analysis of some information types such as word order. it can also be used to do some limited participant tracking, but for texts with a large number of referents it is not practical.

5.3 FILTER. This feature is one of the most powerful features of *Toolbox*. This feature allows the user to delimit a particular database to include and/or exclude the exact information desired to be viewed. For example, if one wants to view only the records of a text which have an SVO word order, then once this is specified in the filter then only those clauses can be viewed and studied. When the filter is turned off, then the entire database is once again viewable. Many other types of filters can be set up that range from simple to complex (including combining information in multiple fields that should be viewed or not viewed). The filtering capabilities are enhanced even more with the large number of choices that the metatagging method introduced here allows. Finally combining the browse mode and the filtering capabilities allows for even greater research capabilities.

Semantic Tier		NP Identification Tier		Grammatical Tier		Topicality Tier		Pragmatic Tier	
A	agent (actor)	PR	pronoun	S	subject	S1-S4	codes used for Levinsohn & Dooley analysis	+	definite
P	patient (undergoer)	NP	noun phrase	O	object	*	dummy symbol	-	indefinite
LOC	locative oblique	P1	pronoun set 1 (absolute case)	Q	oblique				
ABL	ablative oblique	N1	noun phrase set 1 (absolute case)	O2	second object				
O2	second object (undergoer)	N2	noun phrase set 2 (genitive case)	I	intransitive subject				
EQ1	first equative noun phrase	N1/RC	noun phrase set 1 as head of relative clause	Pred	predicate of a verbless clause				
EQ2	second equative noun phrase	RC	relative clause	Qpred	oblique predicate of a verbless clause				
Narg	non-predicate noun phrase argument for verbless clause	ZR	zero anaphora	Left-d	left-dislocated noun phrase				
EX1	first existential clause's noun phrase	P2	pronoun set 2 (genitive case)	APPOS	appositional noun phrase				
EX2	second existential clause's noun phrase	P3	pronoun set 3 (inverse prefix, genitive case)						
Exarg	existential clause's argument	P1/RC	pronoun set 1 as head of relative clause						
C	complement clause functioning as a clausal argument								
APPOS	appositional noun phrase								

FIGURE 6. Examples of abbreviations used for annotations in Pendau participant tracking

6. EXAMPLES FROM PENDAU. This section lists several sample records according to various grammatical categories drawn from the endangered Pendau language (see Quick 2003, 2007). These are representative examples of various kinds of clauses or other categories. They are only illustrative here and not necessarily definitive in how the metatagging approach may be used to annotate texts. Also note that typically there are separate interlinearized lines for part of speech (\ps) and the gloss (\ge). I have merged these two lines into the gloss line (\ge) to simplify the examples for the presentation. Figure 6 lists the abbreviations used in the Pendau participant tracking fields.

6.1 PLURAL NPs. Example (4) illustrates two different cases for handling plural referents. The plural pronoun is used for two referents in this story about the monkey and the turtle. These are coded in separate fields as participant one and participant two, but have the exact same annotation. They are both actors and the grammatical subject of the clause. They are also marked as definite as they have already been introduced prior to this clause (as is typical of pronouns). The third referent 'fish' could be singular or plural just as it is in English. The approach I generally take is not to indicate plural referents as multiple referents in the annotation unless they are unpackaged later in the text as distinct referents. So in

this text the 'fish' searched for in this first instance is indicated as an indefinite noun phrase and is the undergoer and grammatical object of the clause. This is a rather common feature of minor 'props' in a text, and there may also be later instances of 'fish' being searched for which are probably not the same fish. The fourth referent annotated in this clause is the oblique introducing the river.

(4) \rf turtle.pin 003b

\pen Jimo	ma'o	nelolo	bau	ribangkalang.
\mr jimo	ma'o	N-pe-lolo	bau	ri=bangkalang
\ge 3PL/AB	go	RE-SF/DY-search_for	fish	LOC=river

\fte They went to search for fish in the river.
\wo SVOQ
\p1 A_*_P1_+_S
\p2 A_*_P1_+_S
\p3 P__N1_-_O
\p4 LOC_*_N1_-_Q
\dt 24/Apr/2000

6.2 INTRANSITIVE CLAUSES. Example (5) illustrates the second half of a coordinate sentence that is a stative clause. The single argument is an undergoer. Grammatical subjects are marked with a capital I (for Intransitive grammatical subject). This is an important distinction in the annotation system in order to distinguish intransitive clauses from transitive clauses in the tabulation of topic continuity statistics. In Pendau the stative subject is marked with the absolute case (i.e. common nouns are unmarked in the absolute case).

(5) \rf Daras_fish_story05.035d

\pen	o	barumbang	noogemo.
\mr	o	barumbang	no-oge=mo
\ge	and	wave	ST/RE-large=COMP

\fte …and the waves were huge.
\wo SV
\it1 event
\p28 P_*_N1_+_I
\dt 11/Oct/2006

6.3 VERBLESS CLAUSES. Example (6) illustrates an equative clause. As with other verbless clauses one argument can be analyzed as the grammatical subject, and the other noun phrase as the predicate. One approach to annotating the word order is simply to designate one noun phrase as number one (e.g. EQ1) and the second noun phrase as number two (e.g. EQ2). Since an equative clause is typically a description it is noted here as nonevent information. The decision made for annotating the word order can now also be used for the participant tracking field. Since the equative clause refers to the same participant it is useful to differentiate in the semantic tier between the two noun phrases. Although one could identify verbless clauses as an 'intransitive clause' (e.g. see Dixon 1988: 63-68), for the metatagging system presented here it is not necessary to identify a semantic role (which in

any case would be simply a 'single argument' (or 'S' as contrasted with 'A' and 'P'). It also may be more often helpful in the various analytical methods available in *Discourse Profiler* to maintain a separation for the coding between verbal clauses and verbless clauses. For the grammatical tier there are probably several possibilities how these can be indicated. In this example I have marked the first equative noun phrase (EQ1) as the subject (S), and the second equative noun phrase (EQ2) as the predicate (Pred).

(6) \rf Daras_fish_story05.028

\pen Bau	tono'uore	uo,	topenyo	repa.
\mr bau	to=no'u-ore	'uo	tope=nyo	repa
\ge fish	RM=1SG.IV/RE-pull	yonder	name=3SG/GE	snapper

\fte The fish that I pulled up there, it's name is a snapper.
\wo EQ1_EQ2
\it1 nonevent
\p9 EQ1_*_N1_+_S
\p9 EQ2_*_N1_-_Pred
\dt 06/Oct/2006

6.4 FLEXIBILITY. The *Discourse Profiler* metatagging system can be very flexible to meet individual needs. For example, Pastika (1999, 2006) intended to use this metatagging system to work on topic continuity analyses of Balinese texts, and instead of using five elements of information in the participant tracking fields, he used four elements. However, since he still closely followed the constraints of the metatagging system, it is still possible to take his data and simply add a dummy character at some consistent point to make up for the deficiency (since Discourse Profiler currently requires five elements in each participant field). He also shows another possibility for adapting the metatagging system in demonstrating in his database that one does not need to label the annotation fields as required by the software, and even this can be deleted and manipulated with a global automatic change in order to make his database work in the current version of Discourse Profiler. This also highlights the possibility of changing Discourse Profiler in a future version (in its second generation perhaps, or sooner) in order to make the program more flexible to allow optional elements in these fields.

7. DISCOURSE PROFILER IN THE FUTURE. The *Discourse Profiler's* metatagging system presented here is offered as a new tool to add to the fast growing inventory of ways to annotate endangered languages' texts. The first advantage is that it leverages the software tool *Toolbox* which many linguists are already using (and for which there are many people who can help new users). The second advantage is that the metatagging system presented here is a multipurpose system. Although the metatagging system was developed along with the development of *Discourse Profiler*, the separation of the annotated text database in *Toolbox* that is stored as a plain text from the proprietary *Discourse Profiler* software fulfills the current criteria for archiving texts.

The software developed for *Discourse Profiler* is still in its first generation, and the metatagging system has not been exhaustively tested. A current drawback of the *Discourse Profiler's* metatagging system is that it is still necessary to enter the metatagged data manu-

ally into *Toolbox*. As has already been mentioned, there is some help from *Toolbox* if one uses the semantic range fields, however this approach is still not the ideal way to enter an annotation. The semantic range fields can be used for both types of fields when needed or helpful.

Since the span analysis fields are highly abstract, there does not seem to be a lot of potential for developing an automated tagging feature useful for all languages. There may be individual information type fields that can be created through the use of macros for certain types of information that could theoretically be drawn from an interlinearized text, but this will likely have to be left up to individual situations.

In the near future I expect to develop an automated tagging feature for the participant tracking fields. The reason there is strong potential for this is based on the fact that texts are structured and that the five tiers of information provide the basis for a rich variety of analyses as demonstrated by the *Discourse Profiler* software. These analyses can theoretically be reversed so to speak, and used as algorithms to build a kind of weak artificial intelligence that allows the tagger to make 'guesses' of the most likely annotations that would be needed for a given participant in a text. Essentially the tagger would provide a short list of the most likely tags for the participants of a clause.

As more languages are documented, the more likely it will be that some of the texts will be annotated with discourse information when there is an easy to learn system that provides a solution for complementing the current archiving goals for endangered languages' texts. I propose that the metatagging system presented here as developed for the *Discourse Profiler* can serve as a current robust solution to the current gap in annotating texts with discourse information. The potential for an automated tagger would of course dramatically increase the capability for paving the way for annotating texts with discourse information more rapidly, allowing for the possibility of increasing the potential number of texts archived for any particular language with these annotations.

The ultimate goal that has been focused on in this paper has been to offer another tool for documenting languages with an additional breadth of information that can be used for the conservation and preservation of endangered languages. There is also the bonus that this system provides the added potential of performing discourse analysis of texts as well as a variety of other linguistic analyses which should make it more attractive to an even wider range of linguists.

Adding the discourse-text annotations can be kept to a minimum load of additional work, and I am proposing that at least some texts of a corpus that will be archived should be annotated with some discourse-text information. Since one of the goals of archiving texts from an endangered language is to provide a robust documentation of a language, I propose that this software and metatagging system provides a method that can contribute to this important task and that it complements current proposals (e.g. Gippert, Himmelmann and Mosel (eds.) 2006).

References

Andrews, Avery. 1985. The major functions of the noun phrase. In *Language typology and syntactic description, Clause structure*, Shopen, ed. , Vol. I. Cambridge: Cambridge University Press.

Carlson, Lynn and Daniel Marcu. 2001. Discourse tagging reference manual. Unpublished manuscript. http://www.isi.edu/~marcu/discourse/tagging-ref-manual.pdf.

Chafe, Wallace L. 1976. Givenness, contrastiveness, definiteness, subject, topics, and point of view. In *Subject and Topic*, Charles N. Li, ed., pp. 25-56. New York: Academic Press.

Chafe, Wallace L. 1987. Cognitive constraints on information flow. In *Coherence and Grounding in Discourse*, Russell S. Tomlin, ed., pp. 21-51. Amsterdam: John Benjamins.

Dixon, R. M. W. 1988. *A grammar of Boumaa Fijiaan*. Chicago: University of Chicago Press.

Dooley, Robert A., and Stephen H. Levinsohn. 2001. *Analyzing discourse*. Dallas: SIL International.

Dryer, Matthew. 1994. The discourse function of the Kutenai inverse. In *Voice and Inversion*. Talmy Givón, ed. Amsterdam: John Benjamins Publishing Company.

Edmondson, Jerold A. and Donald A. Burquest. 1992. [1998, third edition] *A survey of linguistic theories*. Dallas: The Summer Institute of Linguistics.

Gippert, Jost, Nikolaus P. Himmelmann and Ulrike Mosel, eds. 2006. *Essentials of language documentation*. Berlin: Mouton de Gruyter.

Gildea, Spike. 1994. Semantic and pragmatic inverse: "Inverse Alignment" and "Inverse Voice" in Carib of Surinam. In *Voice and Inversion*, Talmy Givón, ed. Amsterdam: John Benjamins Publishing Company.

Givón, Talmy. 1994. The pragmatics of de-transitive voice: Functional and typological aspects of inversion. In *Voice and Inversion*, Talmy Givón, ed. Amsterdam: John Benjamins Publishing Company.

Givón, Talmy. 1990. *Syntax a functional-typological introduction. Vol. II*. Amsterdam: John Benjamins.

Givón, Talmy, ed. 1983. *Topic Continuity in Discourse: A Quantitative Cross-language Study*. Amsterdam: John Benjamins Publishing Company.

Givón, Talmy, ed. 1994. *Voice and Inversion*. Amsterdam: John Benjamins Publishing Company.

Grimes, Joseph E. 1975. *The Thread of Discourse*. The Hague: Mouton.

Himmelmann, Nikolaus P. 2006. The challenges of segmenting spoken language. In *Essentials of Language Documentation*, Jost Gippert, Nikolaus P. Himmelmann, and Ulrike Mosel, eds. Berlin: Mouton de Gruyter.

Hopper, Paul J. and S. A. Thompson. 1980. Transitivity in grammar and discourse. *Language* 56.251-99.

Longacre, Robert E. 1983. *The grammar of discourse*. New York: Plenum Press.

Longacre, Robert E. 1989. Two hypotheses regarding text generation and analysis. *Discourse Processes* 12: 413-460.

LONGACRE, ROBERT E. 1996. *The grammar of discourse*. Second edition. New York: Plenum Press.

PASTIKA, I WAYAN. 1999. Voice selection in Balinese narrative discourse. PhD diss., Australian National University.

PASTIKA, I WAYAN. 2006. Voice selection in Balinese narrative discourse. Denpasar, Bali: Pustaka Larasan.

QUICK, PHIL. 1977. Her father's store. Unpublished fiction short story.

QUICK, PHIL. 1996. Multilinear discourse analysis software demonstration. In H. Andrew Black, Alan Buseman, David Payne, Gary F. Simons (eds.), *Proceedings of the 1996 General CARLA Conference, November 14-15, 1996, pp. 291-309*. Waxhaw, NC/Dallas: JAARS and Summer Institute of Linguistics.

QUICK, PHIL. 1997. Active and inverse voice selection criteria in Pendau, a Western Austronesian language. In *Proceedings of the Seventh International Conference on Austronesian Linguistics*, pp. 461-482. Amsterdam/Atlanta: Editions Rodopi B.V.

QUICK, PHIL. 2002. A sketch of the primary transitive verbs in Pendau. In *The historical and typological development of Western Austronesian voice systems*, ed. by Fay Wouk and Malcolm Ross. Canberra: Pacific Linguistics.

QUICK, PHIL. 2003. A grammar of the Pendau language. PhD diss., Australian National University.

QUICK, PHIL. 2005. Topic continuity, voice, and word order in Pendau. In *The Many Faces of Austronesian Voice Systems: Some New Empirical Studies*, I Wayan Arka and Malcolm Ross, eds., pp. 221-242. Canberra: Pacific Linguistics.

QUICK, PHIL. 2007. A grammar of the Pendau language of Central Sulawesi, Indonesia. Canberra: Pacific Linguistics.

SCHULTZE-BERNDT, EVA. 2006. Linguistic annotation. In *Essentials of Language Documentation*, Jost Gippert, Nikolaus P. Himmelmann, and Ulrike Mosel, eds. Berlin: Mouton de Gruyter.

TABOADA, MAITE and WILLIAM C. MANN. 2006. Rhetorical Structure Theory: Looking back and moving ahead. Unpublished manuscript. http://www.sfu.ca/rst/pdfs/Taboada_Mann_RST_Part1.pdf

Phil Quick

phil.quick@sil.org