Conversational Analysis Coding Guidelines For the English in the Pacific Northwest Research Study

(Based upon the MUC-7 NE Simple Named Entity Guidelines)

author: Alicia Beckford Wassink (wassink@uw.edu), Department of Linguistics, University of Washington, Seattle, Washington, 98195-4340

1 Introduction

This document outlines and illustrates a coding scheme for sociolinguistic data. The goal was to provide a simple scheme to allow for systematic detection, tagging, and search of a minimal, but broad range of topics of interest to researchers in a range of disciplines who work with textual or audio-recorded and transcribed data, such as is commonly done by researchers in the fields of variationist sociolinguistics, ethnography of speaking, discourse analysis, and linguistic anthropology to name a few.

Sociolinguists generate voluminous amounts of audio-recorded conversational speech in the process of data collection. Sometimes only small amounts of these audio-recorded data are ever transcribed or analyzed (e.g., sometimes a study requires analysis of only a few phones, exemplifying variants of sociolinguistic variables). We think these recordings and transcriptions are valuable, but their value is rarely fully realized. We may increase their value by rendering these materials repurposable for other projects (by using a standardized set of tags that are transparent to most practitioners rather than customized tags understood only to the analyst), or by providing documentation of the coding scheme as a means of easily expanding the amounts of transcribed material. This would make these data more valuable to researchers and to the field, more generally.

In the PNWE project, we use a minimal set of tags to mark a range of topic types that are volunteered in spoken conversation, from personal identity claims, to language ideology, to settlement history. Specific projects may require additional, more specific, project-level subtags to further encode topics of interest to a particular study. A further goal is to allow, for variationist sociolinguistics, the ability to make transparent comments that speakers volunteer about language varieties (dialects, sociolects, etc.), speakers of these varieties, specific linguistic forms, speech communities, and notions about community membership and identity.

This last goal, of making the content of speakers' comments more transparent (or "findable"), highlights a particular value of this scheme, in supporting thematic content analysis. Stemler (2001) defines content analysis as "a systematic, replicable technique for compressing many words of text into fewer content categories based on explicit rules of coding". Essentially, content analysis is a set of complex techniques used in qualitative research, allowing for compression of large amounts of data into a focused set of themes[1]. Themes that emerge from the process may be used in theory-building. In the *emergent coding* approach to accomplishing content analysis, categories are not supplied by a theory, but are established after preliminary analysis of the data, and over iterative passes (sometimes by multiple analysts) (Smith, 2000). Sociolinguists often look to volunteered comments to allow speakers to speak for themselves, to tell the analyst about the frames of reference through which view the world,

and ascribe social meanings. Language allows us to make inferences about subjective experience, interpret reactions, associate cultural conventions with linguistic practices. The careful, systematic coding of transcribed data allows the analyst to perform such content analyses. We use content analysis in the PNWE project to aid in uncovering themes in the recorded data: to facilitate our understanding of how speakers view the linguistic landscape(s) of the Pacific Northwest, construct notions of community in the PNWE, and view linkages between ethnicity and language variation.

2 Coding Methodology Overview

2.1 Coding Scheme

Tag level: utterance or utterance set (e.g., multiple adjacent utterances, such as where a narrative extends over several turns produced by several speakers).

Tag method: inline or standoff

Usage: Tags are 3-letter codes, e.g., [IDE], that associate a theme with an utterance.

An utterance or utterance-set may be tagged with more than one 3-letter code, where a good argument may be made that multiple tags apply.

Note that annotations **may not** overlap or embed in the text. In other words, every annotation must end before another can begin.

Interloperability with other annotation schemes: Codes are intended to supplement the OLAC:discourse-type tagset. Tags are designed to be used in conjunction with extensions to the OLAC controlled vocabulary:

- OLAC:language
- OLAC:field
- OLAC:data-type
- OLAC:participant

May be used in conjunction with other systems that represent discourse-acts (e.g., SWB-DAMSL, MRDA) that encode speech act phenomena.

Resource-level metadata tags are not included.

2.2 ELAN

The methods described in these pages are currently implemented in a tier-based annotation system, using ELAN annotation software (http://tla.mpi.nl/tools/tla-tools/elan/). For this project, 3-letter theme codes are supplied on a single tier called THEMES that is created for each transcribed conversations. Theme units are time-aligned with the orthographic transcription tier in an .eaf file.

Below is a screenshot illustrating ordering of tiers:

[[Figure here—ELAN screenshot]]

3 Coding Categories

3.1 What gets coded

For convenience, the coding categories described below are displayed visually in Appendix A.

The overarching goal of the PNWE Conversational analysis guidelines project was to allow tagging of PNWE conversation transcripts in a way that would make them maximally useful to other researchers. We recognize the Dublin Core Metadata Initiative and Open Language Archive Community (which created extensions to the DCMI element set) as bodies that are setting standards for transparent description, cataloguing, repurposing and extensibility of language resources.

The codes below are intended for compliance with Level-1 standards set forth in the Dublin Core Metadata Initiative (DCMI). Correspondences between DCMI types and UW Sociolinguistics Laboratory current practices are described in the document `MetadataCorrespondences.htm`. The DCMI standards were designed primarily for resource-level metadata specification (recordings, images). The three Level-1 DCMI categories dc:Subject, dc:Type, and dc:Language provide for a small amount of top-level summarization of resource content. However, these categories are insufficient for true conversation analysis:

> DC:Subject - The topic of the resource.
> Typically, the subject will be represented using keywords, key phrases, or classification codes. Recommended best practice is to use a controlled vocabulary. To describe the spatial or temporal topic of the resource, use the Coverage element.

DC:Type - The nature or genre of the resource. (collection, dataset, event, image, interactive resource, sound, text, stillimage, etc.)

Recommended best practice is to use a controlled vocabulary such as the DCMI Type Vocabulary [DCMITYPE]. To describe the file format, physical medium, or dimensions of the resource, use the Format element.

DC:Language - A language of the resource.
from RFC4646:

Language tags are used to help identify languages, whether spoken, written, signed, or otherwise signaled, for the purpose of communication. This includes constructed and artificial languages, but excludes languages not intended primarily for human communication, such as programming languages.

http://www.ietf.org/rfc/rfc4646.txt

THEME FAMILY: General, corpus-neutral category or topic.

THEME: Specific instantiation of one of the theme families from the researcher's corpus or text, making use of project-specific labels.

QUOTATION: illustrative transcribed content or text.

3.2 Theme Families (overview)

Types of themes that emerge in running speech have been grouped into the following broad categories or THEME FAMILIES:

Individual Identity Claims [IND]: Self-identifying statements. Subject claims an identity for self, or associates self with a group.

Group-Level Identity Claims [GIC]: In-group identifying statements. Subject claims an identity as part of a collective, or asserts group membership for self by claiming group traits for him/herself.

Named Entities (several codes): MUC-7 named entity codeset is used. An entity is some object in the world -- for instance, a place or person. A named entity is a phrase that uniquely refers to an entity by its proper name, acronym, nickname or abbreviation. Includes organizations [ORG], Personal names [PER], Title/Role [TTL], Locations [LOC]. We add to the MUC-7 codeset two new labels,

Language Variety [LAN], and Community/Ethnic Group [GRP].

Community Descriptions [COM]: Commentary regarding groups and subcultures about which the subject asserts knowledge. Will often follow a Named Entity tag, but not always. This tag is for use when subject describes a community's features, provides characterizations, or lists social network membership criteria (e.g., notions about what makes the community unique).

Narratives or Stories [NAR]: Stories or events told by the subject. May include stories the subject retells that they heard from others (e.g., passed down within their families). Examples include migration or settlement history, family history, ritual stories.

Ideological Alignment Statements [IDE]: Value-evidencing statements or judgments. Relate information about subject's beliefs about world-view, education, culture, social groups, particular individuals or events, etc. Do not use for the telling of stories (see [NAR], above). Do not use for ideological statements related to language (see Language Awareness Statements, below).

Language Awareness Statements (several codes): This category is for language-related statements. Language Attitudes and Use Statements [ATT] is used for tagging subjective statements about language users or varieties. Neutral Language Awareness Statements [NLA] is used for tagging volunteered illustrative linguistic forms intended to "illustrate" how people or groups talk in a way that does not have clear positive or negative overtones, or when the subject appears to be offering a neutral description of something they have "heard".

Inter-Group Contact Statements [IGC]: Stories told or statements made by the subject that include reference to two or more groups coming into contact. May include contact between any combination of language varieties, community or ethnic groups, and organizations. Examples include “us vs. them” scenarios, interactions with cultures not native to the speaker, and relationships between organizations and communities.

Examples of each theme family are provided in §3.3, below.

3.3 Theme Families

In this method, the most general way of registering “what is talked about” in a conversation is to tag a THEME FAMILY. THEME FAMILIES are general topic types that might occur across a range of corpora. They are “things people talk about” in the most general sense possible. They may be people, places, things or ideas.

Each THEME FAMILY comprises 1 or more THEMES. Particular conversations talk about **particular** people, places, things, or ideas. The sections below describe each theme family in further detail, with illustrative examples and associated themes, and guidance for selecting among multiple candidate tags.

3.3.1 Individual Identity Claims [IND]

Self-identifying statements. Subject claims an identity for self, using first person singular forms ("I am").

```
I am crazy excited about Obama's election.

I'm a straight-shooter. I tell it like it is.
```

3.3.2 Group-Level Identity Claims [GIC]

In-group identifying statements. Subject claims an identity as part of a collective,

```
We are the Yakama.
```

or asserts group membership for self by claiming group traits for him/herself,

```
I am a diehard liberal.

I think – I think uh – basically – all of our ancestors
were pretty much pioneers and settlers and – looking for
something – different and I think – I think that kind of
personifies – people – here – they feel like – we're kind
of up in this – corner – you know kind of away from
everything…
```

A more specific theme may unify a group of [GIC] comments, such as claims that relate to being an authentic member of the group, or articulate the speaker's sense of being from the neighborhood. In this case we may want to name a project-level THEME to use together with the 3-letter THEME FAMILY code, allowing for more specific searching and sorting of themes than is possible using THEME FAMILY alone:

```
[GIC: Sense of being from Yesler Terrace]
```

3.3.3 Named Entities

Here, we borrow from the MUC-7 named entity codeset. **Linguists don't always want to know each and every person, place or thing that was named in a conversation. But sometimes this is useful. Linguists do often want to**

know what languages or language varieties were named and/or used in a linguistic dataset. This coding scheme allows tagging of a range of types of entities of interest to linguistic analysis projects: organizations [ORG], Personal names [PER], Title/Role [TTL], Locations [LOC]. We add to the MUC-7 codeset two new labels, Language Variety [LAN], and Community/Ethnic Group [GRP]. Examples of MUC-7 categories are reproduced directly from the file: `SimpleNamedEntityGuidelinesV6.4.pdf`

3.3.3.1 Organization [ORG]

Organization entities are limited to corporations, institutions, government agencies and other groups of people defined by an established organizational structure.

General entity mentions such as "the police" and "the government" should not be tagged, since these are not unique proper name references to specific entities. Tag all proper name mentions of groups with a defined organizational structure. These include:

Businesses
`[Bridgestone Sports Co.] profits`
Stock exchanges
`[NASDAQ] shares`
Multinational organizations
`[European Union] representatives`
Political parties
`[GOP] hopeful`
Non-generic government entities
`[the State Department]`
Sports teams
`[the Phillies]`
Military groups
`[the Tamil Tigers]`

Many other kinds of entities refer to facilities or buildings that are primarily defined by their established organizational structure, and can do things like issue statements, make decisions, hire people, raise money and so on. A mention of such an entity should be tagged as an ORGANIZATION when it functions like an ORG in the document. These include things like:

Churches and other religious institutions
`[Trinity Lutheran Church]`
Hospitals
`[Finger Lakes Area Hospital Corp.]`
Hotels
`[Four Seasons Hotel Group]`
Museums
`[the Guggenheim Museum]`
Universities
`[the University of Chicago]`

Government offices
```
[the White House]
```
Note that definite and indefinite determiners ‘the’ and ‘a’ are included in the annotation, except for cases when they quantify something other than the tagged entity, as in the following examples:
```
A [Gulshan Hotel] spokesman the [U.S.] Vice President
```
As in the above examples, this exception is particularly common when the tagged name is used in the pre-modifier (adjective) position.

3.3.3.2 Person Names [PER]

People may be specified by name, nickname or alias. Family names should also be tagged as PERSON. Names of deceased people, as well as fictional human characters appearing in movies, television, books and so on, should be tagged as PERSON entities. Religious deities should also be tagged as persons.

Other types of named entities like animals, inanimate objects and monetary units will not be annotated.

3.3.3.3. Location Names [LOC]

Location entities include names of politically or geographically defined places (cities, provinces, countries, international regions, bodies of water, mountains, etc.). Locations also include man-made structures like airports, highways, streets, factories and monuments.

mountains, fictional or mythical locations, and monumental structures, such as the Eiffel Tower and Washington Monument. For instance:
```
the collapse of the newly-constructed [Teton Dam]
LOC

the dispute over votes in [Dade County]
LOC

[The Walt Whitman Bridge] remained closed
LOC

repairs began on a 10-mile stretch of [the Alaskan Pipeline]
LOC

[The Garden State] is known for its tomatoes.
LOC
```

3.3.3.4 Titles, Roles and Appositives [TTL]

(Note: We find that this MUC-7 category can be difficult to tag separately from [PER]. The MUC-7 guideline is to include the [TTL] + [PER] for the sequence Title + Name. Titles may not be of direct interest in this research project. For this reason, it may be acceptable to simply code [PER] and not separately code [TTL].) But here’s the general MUC-7 guideline:

Titles, roles and honorifics such as "Mr." and "President" are tagged as title entities and are separated from the individual's name. For instance, in the following sentence, there are two separate entities marked:

```
[Vice President] [Cheney] visited the site.
```

For this task we define titles and roles as occurring either directly before or directly after a person name. Therefore, titles and roles are only tagged when they occur directly next to the person name they modify. In the following example, for instance, the phrase "Vice President" is not considered a title and is not tagged:

```
The strongest supporter was the Vice President.
```

If a title contains within it a taggable entity, tag that entity separately. For instance:

```
[Microsoft] [Chairman] [Bill Gates] stated that...
ORG  TTL  PER
```

You may occasionally encounter an appositive like "Jr.", "Sr.", and "III". These are considered part of a person name and should be marked as part of the name, for instance:

```
[Mr.] [Albert Franklin, Jr.] was part of the research team.
TTL PER
```

Finally, sometimes the name of the person is split into two pieces by the title. In these cases, we will annotate the two pieces of the PERSON name as two separate PERSON entities:

```
[Alfred] [Lord] [Tennyson]
PER TTL PER
```

Some more examples of names and titles:

```
[GlobalCorp] [Vice President] [John Smith]
ORG  TTL  PER

[Treasury] [Secretary] [Jackson]
ORG  TTL  PER

the [U.S.] [Vice President], [Dick Cheney]
LOC TTL  PER

[Justice] [Minister] [Giovanni Maria Flick]
ORG  TTL  PER

[British] [Rashtrodut] [Anwar Coudhury]
LOC  TTL  PER

[Mission Control] [Chief] [Vladimir Solovyov]
```

```
ORG TTL PER
```

3.3.3.5 Annotating codeswitches

We note that if the speakers codeswitch between linguistic varieties, this code may be used to reflect such switches.

```
I begin in English y termino en español
[LAN: English~Spanish]
```

3.3.4 Community Descriptions [COM]

Commentary regarding groups and subcultures about which the subject asserts knowledge. Will often follow a Named Entity tag, but not always. This tag is for use when subject describes a community's features, provides characterizations, or lists social network membership criteria (e.g., notions about what makes the community unique).

We are particularly interested in commentary regarding groups and subcultures (e.g., Jets, Thunderbirds, Cobras, Lames), network or group membership criteria, statements about liminal members (who is named as an outgroup member), naming exercises [[LD: do we have an example of this?]]

```
Yeah, but – boy – I think Seattle women embraced – pants
and low heeled shoes really fast…
```

We may also identify THEMES for our corpus below the level of the THEME FAMILY [COM]:

```
[COM: Yesler Terrace as a Landing Site]
```

3.3.5 Narratives or Stories [NAR]

[NAR] encodes stories or narratives. Often we are interested in Migration stories, settlement history, family history, as well as relating of culturally-ritualized stories.

This tag is typically used across multiple utterances, sometimes across speakers (what above was referred to as an utterance-set).

3.3.6 Ideological Alignment Statements [IDE]

[IDE] encodes statements of an evaluative nature; value judgments, belief-evidencing statements (about world-view, citizenship, family values), prescriptivist statements (that name a held, positively viewed cultural value, such as what education accomplishes, or what language should be like, or the

boundaries of appropriate linguistic or social behavior). As this last sentence should make evident, standard-language related notions are tagged as [IDE] statements.

Even though we were poor, she taught us to speak proper.

I tell you wh- the thing that bugs me the most is the difference in- outlook and mindset – because Seattle used to be a modest place – and people were – people di- were not flashy and showy – I mean I was probably thirty five before I ever saw a limousine…

THEMES below this THEME FAMILY might include:

[IDE: melting pot]

3.3.7 Language Awareness Statements

This category further divides into two subcategories for differentiating between types of language-related statements. Language Attitudes and Use Statements [ATT] is used for tagging subjective statements about language users or varieties. Neutral Language Awareness Statements [NLA] is used for tagging volunteered illustrative linguistic forms intended to "illustrate" how people or groups talk in a way that does not have clear positive or negative overtones, or when the subject appears to be offering a neutral description of something they have "heard".

3.3.7.1 [ATT] Language Attitudes and Use Statements

This category includes evaluative statements about language users or specific linguistic varieties. Here, also, we tag indexical statements linking linguistic forms to groups.

They sound uneducated.

Hick

They use that hard-g.

They say the ladies' name Roof instead of Ruth.

3.3.7.2 [NLA] Neutral Language Awareness Statements

This category includes volunteered illustrations of phonetic, morphological and syntactic forms that is NOT evaluative or otherwise value-evidencing, as well as apparently neutral descriptions of linguistic varieties.

And there's a different – there's a different staccato or a

– a – I don't know if XXX if I'm saying the right word but – but there's – um – the Californians talk – more like us maybe less intensely or something…

Well, we drink 'pop' here…

I would – I would say that probably just – you know thinking off the top of my head probably – of course this isn't scientific but um that – I think probably people – northwest Washington – um – probably use slang a lot more – that – it's not the real proper English like you might here in – other places – certainly not like England.

It feels more sing songy in Georgia…

3.3.8 Inter-Group Contact Statements [IGC]

[IGC] encodes stories told or statements made by the subject; specifically, ones that include reference to two or more groups coming into contact. Often seen in conjunction with a Narrative tag, but not always.

This tag may be used across multiple utterances, and sometimes across speakers (what above was referred to as an utterance-set).

3.4 Specific Themes

Following is a list of specific theme subtags currently in use on the PNWE project. Each of these is coded in the format [XXX: Subtag].

3.4.1 Named Entities

3.4.1.1 [ORG] Organization

Boeing
Sounders
Military

3.4.1.2 [PER] Person

J.P. Patches

3.4.1.3 [LOC] Location

Seattle
Bainbridge
Spokane

Japan

3.4.1.4 [LAN] Language

Japanese
English

3.4.1.5 [GRP] Community/Ethnic Group

Japanese
American
Japanese American
No-no boys
Nisei

4 Caveats and Special Cases

4.1 Caveats: What Not to code

4.2 Special Cases

5 References

Dublin Core Metadata Initiative (2012) http://dublincore.org/documents/dces/

Krippendorf, Klaus (2004). Content analysis: An introduction to its methodology. 2nd ed. Thousand Oaks, CA: Sage Publications.

Open Language Archive Community (2012) http://www.language-archives.org/OLAC/metadata.html

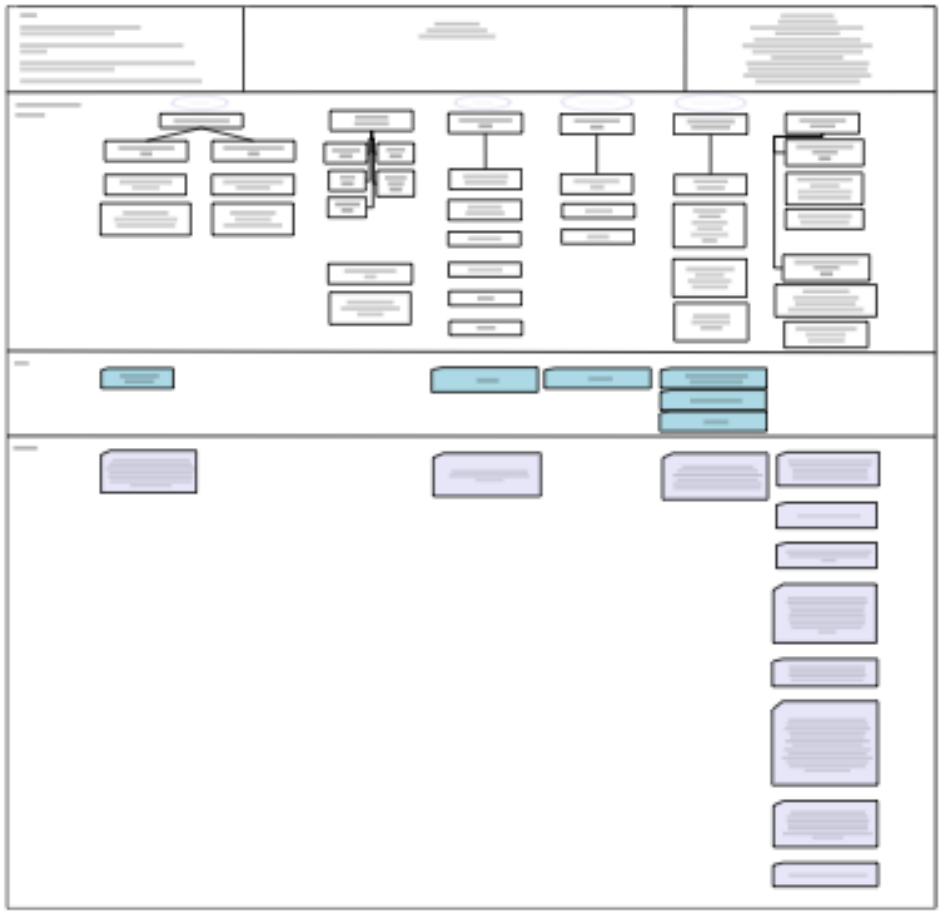
Smith, Charles (2000) Content Analysis and Narrative Analysis, ch. 12. Motivation and personality: Handbook of thematic content analysis. New York: Cambridge University Press.

Stemler, Steve (2001). An overview of content analysis. Practical Assessment, Research & Evaluation, 7(17). Retrieved October 30, 2005 from http://PAREonline.net/getvn.asp?v=7&n=17 .
accessed from:
http://szekedi.uw.hu/ad_7/overview%20of%20content%20analysis.pdf

UW Sociolinguistics Laboratory website (2012)
http://depts.washington.edu/sociolab/resources.htm

6 Appendices

metadataflowchart-ABWrev5.pdf



[1] Krippendorf (2004) identifies five key processes inherent to content analysis:

1. Unitizing. The researcher must establish the unit of analysis (word, meaning, sentence, paragraph, article, news clip, document, etc.).

2. Sampling. Usually the universe of interest is too large to study the content of all units of analysis, and instead units must be sampled. Sampling involves counting, which may require the researcher to develop thesauruses (so different terms with like meanings will be counted under the same construct) and expert systems or other rule engines (so the proper contextual valence is assigned to each counted construct).

3. Reducing. Content data must be reduced in complexity, usually by employing conventional summary statistical measures. Coding and statistical analysis is covered by Hodson (1999).

4. Inferring. Contextual phenomena must be analyzed to provide the context for findings.

5. Narrating. Conclusions in the content analytic tradition are usually communicated using narrative traditions and discursive conventions.