The PURSUIT Corpus: Documentation

Nate Blaylock
Institute for Human and Machine Cognition
Pensacola, Florida, USA

December 2009

1 Introduction

This document describes the PURSUIT Corpus, which was collected at the Institute for Human and Machine Cognition in January and February 2008. It consists of 13 audio and GPS recordings of realtime route descriptions as they were driven in cars in the downtown Pensacola, Florida area. The audio signals have been transcribed and separated into utterances. Additionally, references to geospatial entities has been annotated in the corpus with the name, address, and latitude and longitude information for each referent.

This corpus was created as part of a project for automated understanding of natural language path descriptions. The GPS recording of the actual path taken, along with the annotated geospatial references serve as a "ground truth" for understanding the path description. A set of statistics about the corpus is shown in Table 1.

In the remainder of this document, we first describe the distribution and then how the corpus was collected. We then describe annotation (including transcription) and then conclude.

2 Files in the Distribution

At a high level, the corpus distribution is divided into two directories. `Data` contains the actual corpus data and `Doc` contains documentation. In `Data`, the file

	Total	Average per Session
Length	3 hours 55 minutes	18 minutes 4 seconds
Utterances	3155	243
Annotated References	1649	127

Table 1: PURSUIT Corpus Statistics

`meta.xml` is a NITE XML description of the corpus contents, and `otherMeta.xml` contains additional metadata about the corpus.

The `Signal` directory contains audio recordings (`.wav`) and GPS tracks (`.gpx`) for each car in each session. The naming scheme is as follows:
`<session>.<car position>`

For example, the file `s1.foll.audio.wav` contains the audio recording for the follow car in session 1.

The `Annotation` directory contains the NITE NXT annotation files, and uses the same naming convention as mentioned above. For each car, there are three annotation files: `words` contains the transcription of the audio; `segments` contains the segmentation into utterances; and `locations` contains the annotation of location references.

3 Corpus Data Collection

The data collection methodology is detailed in [BA08]. For convenience, we summarize in this section.

3.1 Setup

Figure 1 shows an example of the data collection setup for the corpus collection. Each session consisted of a lead car and a follow car in downtown Pensacola, Florida. The driver of the lead car was instructed to drive wherever he wanted for an approximate amount of time (around 15 minutes). The driver of the follow car was instructed to follow the lead car. One person in the lead car (usually a passenger) and one person in the follow car (usually the driver) were given close-speaking headset microphones and instructed to describe, during the ride, where the lead car was going, as if they were speaking to someone in a remote location who was trying to follow the car on a map. The speakers were also instructed to try

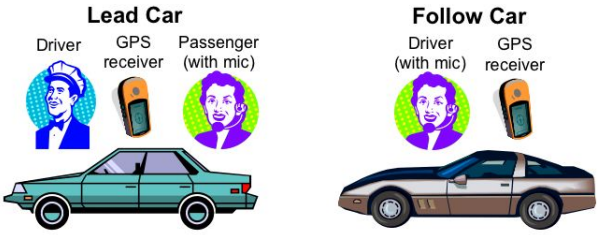

Figure 1: Data Collection Setup

to be verbose, and that they did not need to restrict themselves to street names—they could use businesses, landmarks, or whatever was natural. Both speakers' speech was recorded during the session. In addition, a GPS receiver was placed in each car and the GPS track was recorded at a high sampling rate.

3.2 Data

The corpus contains 13[1] audio recordings of seven paths along with two corresponding GPS tracks from the cars. The average session length was just over 18 minutes. Some sample utterances from the corpus are given below:

- *...and we're going under the I-110 overpass I believe and the Civic Center is on the right side on the corner of Alcaniz and East Gregory Street where we are going to be taking a left turn...*
- *... he's going to turn left right here by the UWF Small Business Development Center heading toward Gulf Power ...*
- *... we've stopped at a red light at Tarragona Street okay we're going now across Tarragona passing the Music House ...*
- *... we're at the intersection of East Gregory and 9th near a restaurant called Carrabas I think and a Shell station just a little south of the railway crossing ...*

[1] In one session only one audio recording was made.

3.3 Synchronization

The resulting audio and GPS track files for each session were synchronized by hand to start and end at the same point in time. As the recording on each device was started separately from the others, this lead to special challenges in synchronization. Using the TESLA annotation and visualization tool for this corpus [BSA09], the annotator adjusted audio and GPS length and starting time by hand until the audio descriptions and GPS tracks seemed to be in concordance.

4 Annotation

The corpus has been manually annotated with transcription, utterance, and location reference information. Before describing these, however, we first describe the annotation format of the corpus.

4.1 Annotation Format

We use the NITE XML Toolkit (NXT) data model [CEHK05] for storing both the corpus and annotations on it. NXT is a general XML data model for multimodal and heavily cross-annotated corpora. In the data model, a corpus is represented as a list of observations, which contain the data for a single session. An observation contains a set of synchronized signals, which are typically audio or video streams associated with the observation, although NXT is broad enough that a signal may be any timestamped stream of data. Annotations are represented as a multi-rooted tree structure, where leaves are segments that are time-aligned with an underlying signal. This allows disparate annotations to be made on and saved with the same corpus.

4.2 Transcription

Transcription of the audio signal was done manually using the Transcriber tool [BGWL00]. The resulting transcription included not only words, but also preliminary utterance breaks that were useful to the transcriber (these were used later to estimate word timing information as discussed below).

Transcription rules were that no punctuation was to be transcribed, except in phrases requiring a hyphen, periods in names with abbreviations, and apostrophes. Proper nouns were capitalized, but the beginnings of utterances were not. Internet

resources such as Google Local were used to verify canonical spellings of proper nouns such as business or street names. Numbered street names were spelled out (e.g., *Seventeenth Avenue*). In cases where the correct transcription could not be determined, the token [unintelligible] was inserted as a word.

The words level in NXT requires not only the list of transcribed words, but also timing information on the start and end time of each word. This was estimated by using the rough transcriber utterance boundaries (described above) for the start and end time of each rough utterance. The start and end time of each individual word were estimated by equally dividing the utterance time into chucks for each word within it.

4.3 Utterance Segmentation

Utterance segmentation was done manually using the TESLA tool. Utterance segments of spoken monologue are admittedly somewhat arbitrary, but annotators were instructed to use cues such as pauses and grammar.

4.4 Location Annotation

References to certain types of locations were segmented and annotated by hand with information about each referent using the TESLA tool. Annotators were instructed to segment the entire referring phrase, as opposed to e.g., just the head-word.

The high-level classes annotated were:

- *Streets*: references to a given street, for example "Garden Street" or "a divided road"
- *Intersections*: references to street intersections, for example "the corner of 9th and Cervantes" or "the next intersection"
- *Addresses*: references to street address, for example "401 East Chase Street" or even "712" (when referring to the address by just the street number)
- *Other Locations*: this class is a grab bag for all other location types that we annotated, consisting of such data as businesses, parks, bridges, bodies of water, etc.

Note that not *all* geospatial entity references have been annotated in PURSUIT—just those that are accessible in our GIS databases. Examples of entities that

	Named	Category	Total
Street	77.2%	22.8%	**48.5%**
Intersection	45.5%	54.5%	**6.8%**
Address	100.0%	0.0%	**0.8%**
Other Loc	67.7%	32.3%	**43.9%**
Total	**71.1%**	**28.9%**	**100%**

Table 2: Breakdown of geospatial entity reference annotations in the PURSUIT Corpus

appear in the corpus but were not annotated are fields, parking lots, sidewalks, railroad tracks, and fire hydrants. These were not annotated only because we did not have access to data about those entities. However, there is nothing inherent in our approach to path understanding which would prohibit the use of those classes of entities, if data were available for them.

Although not all *classes* of entities were annotated, within those classes that were annotated, *all* references to entities of interested were annotated, whether or not they were named. Thus "Garden Street", "a divided road", or even "it" were annotated if they referred to a geospatial entity of interest.

Annotations are also marked with whether an entity reference was *named* (i.e., contained at least part of the proper name of the entity, such as "the Music House" and "the intersection at Cervantes") or *category* (description did not include a name, such as "the street", "a Mexican restaurant", and "it").

All entity references of interest were minimally annotated with their canonical name and a lat/lon coordinate. Streets were annotated with the lat/lon of the street segment from the database closest to the speaker's current location. Where applicable, entities were also annotated with a street address. In cases where the entity was not in the databases, the human annotator searched for the missing data by hand using various resources.

In total, 1649 geospatial entity references were annotated in the corpus. The breakdown of categories is shown in Table 2.

4.4.1 Source Database Information

As noted above, several sources were used to search for geospatial entity information for annotation. The data sources are also noted in the annotation on each reference under the `src` attribute. The two main data sources used are TerraFly

and Google Local (which we will describe in more detail below). Additionally, for references which were not found in either of these databases, the annotator used any of a variety of other methods to find the referent information, including web searches, and even, in a few cases, physically travelling to the location.

The `src` attribute records anything found outside of the two main databases as `human`. If a referent was found in one or both of the databases, the source attribute is listed as `terrafly:<SUB>` or `googleLocal`. In the former, the <SUB> attribute is replaced by the sub-database of TerraFly that it was found in (described below). Also, if a referent was found in both databases, the source includes both sources separated by a semicolon (;).

We now describe the two databases.

TerraFly A primary source used was a custom-made subset of the TerraFly GIS database [RGSG05]. The custom database was made by compiling data from the datasets listed in Table 3.

Google Local Google Local[2] (also known as Google Maps) provides a service for searching for businesses near a location. This dataset does not contain several types of information, including streets, intersections, and bodies of water.

5 Conclusion

This document has described the PURSUIT Corpus, which contains 13 audio recordings and corresponding GPS tracks of realtime path descriptions as the paths were driven in cars. Additionally, location references in the corpus have been manually annotated with geospatial entity information.

Acknowledgements

We would like to thank the following: James Allen, who gave scientific oversight to this corpus development; Bradley Swain, who helped with the annotation and development with the TESLA annotation tool; and Dawn Miller, who also helped with the annotation.

[2] `http://maps.google.com`

Dataset	Code	Description
Street Blocks	*austreets*	Derived from v3.2 (April 1, 2009) NAVTEQ NAVSTREETS Street Data
Intersections	*auintersections*	Derived from v3.2 (April 1, 2009) NAVTEQ NAVSTREETS Street Data
Restaurants	*nv_restrnts*	From v3.2 (April 1, 2009) NAVTEQ POI Data: Restaurants
Public Schools	*public_schools*	2007 National Center for Education Statistics (NCES) Public Schools
Private Schools	*private_schools*	2007 National Center for Education Statistics (NCES) Private Schools
Yellow Pages	*ypages*	2005 Yellow Pages Business Information
Hotels	*hotels2*	Hotels data obtained by merging of data periodically extracted from major reservations systems
Business	*nypages*	Infot, Inc. 2007 Business Database of USA
Travel	*nv_travdest*	From v3.2 (April 1, 2009) NAVTEQ POI Data: Travel Destinations and Facilities
Auto	*nv_autosvc*	From v3.2 (April 1, 2009) NAVTEQ POI Data: Automotive services, gas stations
Recreation	*nv_parkrec*	From v3.2 (April 1, 2009) NAVTEQ POI Data: Recreation facilities
GNIS	*gnis2008_addr*	USGS 2008 Geographic Names Information System (GNIS) Database, supplemented with other sources
Pincorp	*pincorp*	Cities and Incorporated Places – Demographic and Socioeconomic Summary Data from US Census 2000

Table 3: Constituent Datasets of the TerraFly database

References

[BA08] Nate Blaylock and James Allen. Real-time path descriptions grounded with GPS tracks: a preliminary report. In *LREC Workshop on Methodologies and Resources for Processing Spatial Language*, pages 25–27, Marrakech, Morocco, May 31 2008.

[BGWL00] Claude Barras, Edouard Geoffrois, Zhibiao Wu, and Mark Liberman. Transcriber: development and use of a tool for assisting speech corpora production. *Speech Communication special issue on Speech Annotation and Corpus Tools*, 33(1–2), January 2000.

[BSA09] Nate Blaylock, Bradley Swain, and James Allen. TESLA: A tool for annotating geospatial language corpora. In *Proceedings North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT) 2009 Conference*, Boulder, Colorado, May 31–June 5 2009.

[CEHK05] Jean Carletta, Stefan Evert, Ulrich Heid, and Jonathan Kilgour. The NITE XML toolkit: data model and query language. *Language Resources and Evaluation Journal*, 39(4):313–334, 2005.

[RGSG05] N. Rishe, M. Gutierrez, A. Selivonenko, and S. Graham. TerraFly: A tool for visualizing and dispensing geospatial data. *Imaging Notes*, 20(2):22–23, 2005.