UNIVERSITY OF CALIFORNIA
SANTA CRUZ

CHARACTER MODELING THROUGH DIALOGUE FOR EXPRESSIVE NATURAL LANGUAGE GENERATION

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

COMPUTER SCIENCE

by

Grace I. Lin

September 2016

The Dissertation of Grace I. Lin
is approved:

Professor Marilyn A. Walker, Chair

Professor Arnav Jhala

Professor Jean E. Fox Tree

Tyrus Miller
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

Character Modeling through Dialogue
for Expressive Natural Language Generation

by

Grace I. Lin

Conversation is an essential component of social behavior, one of the primary means by which humans express emotions, moods, attitudes, and personality. Conversation is also critical to storytelling, where key information is often revealed by what a character says, how s/he says it, and how s/he reacts to what other characters say.

Interactive narrative systems (INS) are a type of playable media whose applications range from simple entertainment to systems for learning, training, and decision making. Many forms of INS involve interactions with virtual human characters. Thus a key technical capability for such systems is the ability to support natural conversational interaction. While most INS use hand-crafted character dialogue to produce high quality utterances, they suffer from problems of portability and scalability, or what has been called the authoring bottleneck. We believe Natural Language Generation (NLG) is part of the solution to alleviate such burden from authors by automatically generating character dialogue.

Here we focus on the issue of character voice. One way to produce believable, dramatic dialogue is to build stylistic models with linguistic features related to NLG decisions. Film/television dialogue are exemplars of many different linguistic styles that were designed to express dramatic characters. Thus we construct a corpus of film/television character dialogue from screenplays and transcripts publicly available from websites such as the Internet Movie Script Database. We

apply content analysis and language modeling techniques to extract relevant linguistic features to build character-based stylistic models. We also apply machine learning techniques to discriminate characters base on available metadata such as genre, year, and director.

This thesis consists of two parts. The first part involves building a basic character model with film dialogue, and then applying the model to an existing expressive NLG engine to generate different character voices. We then evaluate the generation experiment with a perceptual study, which suggests several natural extensions.

The second part involves building a more refined model with television dialogue in order to explore a broader range of stylistic features that can be used to express dramatic characters. We test the model-fit of character models in two ways: 1) ranking experiments to pick out corresponding character's utterances from a pool of mixed, original characters utterances, and 2) a second generation experiment to test user perceptions of characters.

Acknowledgments

I would like to thank my advisor Professor Marilyn Walker and my committee members Professor Arnav Jhala and Professor Jean E. Fox Tree for their guidance and support throughout the years. I would also like to thank Professor Noah Wardrip-Fruin for being part of my advancement committee and providing advice during the earlier years of my research.

To my family for their love and support. To all my friends and acquaintances, past and present, for helping me and enriching my life at different times of my PhD career. To Jen and Rob for going through the journey together. To members of the NLDS lab for being awesome. To PhD Comic and XKCD for keeping my life sane.

Chapter 1

Introduction

Conversation is an essential component of social behavior, one of the primary means by which humans express emotions, moods, attitudes and personality. Conversation is also critical to storytelling. Key parts of a narrative may be revealed with what a character says, how he says it, and how he reacts to what other characters say.

Interactive narrative systems (INS) are a new type of playable media whose applications range from simple entertainment to systems for learning, training, and decision making [Rowe et al., 2008, Mott and Lester, 2006, Traum et al., 2007, Riedl and Young, 2004, Shaffer et al., 2005]. Many forms of INS involve interacting with virtual dramatic characters. Thus a key technical capability for such systems is the ability to support natural conversational interaction.

In most INS to date, **character dialogue is highly hand-crafted**. Although this approach offers total authorial control and produces high quality utterances, it suffers from problems of portability and scalability [Walker and Rambow, 2002], or what has been called the **authoring bottleneck** [Mateas, 2007].

For example, in the interactive drama action-adventure psychological thriller video game *Heavy Rain*, the player's decisions and actions affect the narrative.

The main characters could be killed, and different choices may lead to different scenes and endings. The script is about 2000 pages long (a movie is about 120 pages), containing 40,000 words of non-linear dialogue, 60 scenes ranging from 15 to 20 minutes, and 15 months in development. In massively multiplayer online (MMO) games like the *Star Wars: Old Republic*, the script was about 40 novels back in 2009 and continues to expand as more quests (story lines) are added.

Even though script writers painstakingly describe different narrative paths in order to maximize users' interactive experience, it is still difficult to achieve full interactivity. For example in *Heavy Rain*, once a character dies, all the scenes associated with the character from that moment on are deleted [Wei and Calvert, 2011]. Wouldn't it be more interesting if you could experience the same scene but with different characters? Different pairs of characters would create different interactive experiences because characters have different personalities, which will be reflected in the style of their dialogue interactions.

We believe that **Natural Language Generation (NLG)** techniques are part of the solution to automatically producing narrative adaptions for interactive stories. This would allow automatically generating scenes that reflect the player's history and choice in story so far, as well as also adapting a scene so that it can be played by different types of characters.

More specifically, we believe that **Expressive Natural Language Generation (ENLG)** offers the potential to address scalability issues and to produce variations in linguistic style that can manifest differences in dramatic characters. It has the potential to produce continuous stylistic variation over multiple stylistic factors by automatically learning a model of the relation between stylistic factors and properties (parameters) of generated utterances [Paiva and Evans, 2004, Paiva and Evans, 2005, Bouayad-Agha et al., 1998].

While we are not the first to have this idea, most of the prior research in this novel area of interactive story and drama generation has focused on how planning mechanisms can be used in order to automatically generate story event structure, as shown in the top portion of Figure 1.1. The figure shows a plan-based representation for how an author goal for the detective to INVESTIGATE would be elaborated into subgoals for FIND CRIME SCENE EVIDENCE, INTERROGATE SUSPECT, etc. Notice that these **representations bottom out in hand-crafted dialogue**. Imagine trying to write narratives for any of 5 different detective characters interacting with any of five villains; the author would be required to create each of the 25 versions of conversation of the same scene.

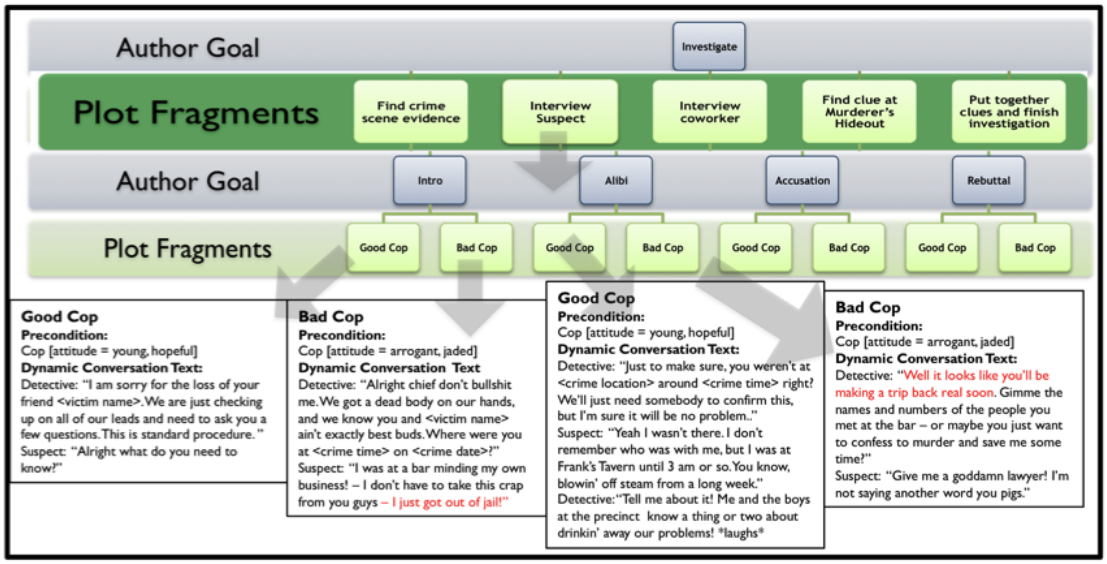


Figure 1.1: Interactive Narrative System - Story Representation

Recent research, including our own, has begun working on generating story dialogue on a turn-by-turn basis [Cavazza and Charles, 2005, Rowe et al., 2008, Lin and Walker, 2011a, Walker et al., 2011a]. While other approaches to dialogue generation have made their contributions to interactive storytelling, only our work to date uses a full scale ENLG architecture for a **full procedural generation of**

dialogue. In addition, we automate the process of creating models of characters from film and TV dialogue and apply learned models to our ENLG engine PERSONAGE [Mairesse and Walker, 2011] and its Python spin-off, PYPER, with new control to generate utterances that can be perceived as having the same personality as the original character, given a completely different story domain [Lin and Walker, 2011a, Walker et al., 2011a, Walker et al., 2011b].

We believe that the stylized, crafted aspects of film/TV dialogue are actually useful for our purposes. Film/TV dialogue is authored deliberately in order to convey the feelings, thoughts and perceptions of the character being portrayed, and the screenplay often specifies the emotion of an utterance with psychological state descriptors. In addition, the dialogue is deliberately constructed to focus the viewer's attention on the character's personality, and the key plot events involving a character and their perceptions, especially in dramatic films as opposed to action. For example, Figure 1.2 shows two different scenes stylistically varied by characters. While both scenes are about getting a ride, Alvy and Annie (*Annie Hall*) have a non-straightforward conversation about it, while The Terminator (*Terminator 2*) only has one line that goes straight to the point.

Topic: Getting a ride	
Film: *Annie Hall* Scene: Lobby of sports club	Film: *Terminator 2* Scene: Biker bar
Alvy: Uh... you-you wanna lift? **Annie:** Oh, why-uh... y-y-you gotta car? **Alvy:** No, um... I was gonna take a cab. **Annie:** [Laughing] Oh, no, I have a car. **Alvy:** You have a car? So... [clears his throat]. I don't understand why... if you have a car, so then-then wh-why did you say "Do you have a car?" ... like you wanted a lift?	**The Terminator:** I need your clothes, your boots & your motorcycle. **Cigar Biker:** You forgot to say please.

Figure 1.2: An Example of Two Different Scenes Stylistically Varied by Characters.

An example of how the same set of utterance can be stylistically varied by

models of film characters, Alvy and Annie from *Annie Hall*, is shown in Table 1.1. The variations are highlighted through different colors. Alvy is neurotic and stutters in his speech. This is expressed through disfluencies such as *err...*, *oh*, and *I see*. The stuttering is expressed through partial-word repetition such as *st-strange*, *br-brought*, and *ge-ge-gentle*. Annie tends to use emphasis and hedging pragmatic markers such as *really*, *sort of*, *I think*, as well as using many tag questions, such as *he isn't my close friend, is he?*

Table 1.1: Utterances Generated using Film Character Models. Utterance variations across models are in different colors or highlights.

Film, Character, and Generated Utterances	
***Annie Hall*: Alvy**	***Annie Hall*: Annie**
I don't know. People say Cartmill is **st-**strange, **alright? Err... on the other hand,** I don't rush to judgment.	**Come on, I don't know, do you?** People say Cartmill is strange **while** I don't rush to **um..** judgment.
Right, I am not sure, would you be? I will tell something you because you br-brought me cabbage.	I don't know. I think that you brought me cabbage, so I will tell something to you, alright?
Oh I am not sure. Wolf wears a hard shell. **On the other hand,** he is **ge-ge-**gentle, **isn't he?**	**Yeah, I am not sure, would you be?** Wolf wears a hard shell but he is **really** gentle.
I see, I don't know. I respect Wolf, wouldn't you? He, however, isn't my close friend.	I see, I am not sure. Obviously, I respect Wolf. However, he isn't my close friend, is he?
Yeah, I don't know. Sparrow conveys excitement to my life, **so** I am **fr-fr-**friends with her.	**Come on, I am not sure. Because** Sparrow brings excitement to my life, I am friends with her, **you see?**

The **hypothesis** of this thesis is that the *authoring bottleneck* can only really be solved with 1) a full scale ENLG, 2) better utilization of narrative structure, and 3) character voice can be learned from dialogue corpora. The overall model describing our approach is shown in Figure 1.3 and summarized in the following paragraphs.

The **Dialogue Corpus** is composed of two raw corpora containing data downloaded from various sources: film scripts from the Internet Movie Database

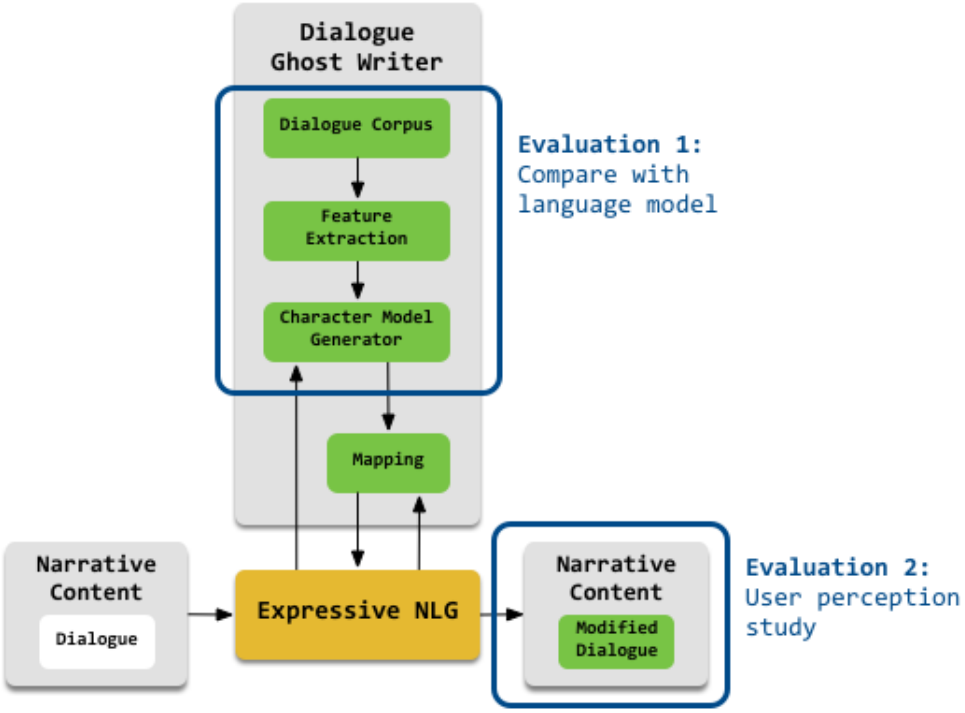


Figure 1.3: Overall Model of Our Approach.

(IMDb) and TV scripts from fan-transcribed website *Big Bang Theory*. These files are pre-processed to extract the dialogue for each character.

We are aware that scripted dialogue is not exactly like spontaneous speech, but this is appropriate for our purpose, since our goal is to produce scripted, stylistic dialogue (for generating stories). We also realize that full character modeling requires the addition of non-verbal cues such as gaze during dialogue with human users [Cassell et al., 1999], which is beyond the scope of our work and therefore will not be addressed.

The **Feature Extraction** component extracts dialogue and acoustic features from the pre-processed data. The main focus is to look for various dialogue patterns to represent how the character talks and to fully utilize the control of 67 stylistic parameters of PERSONAGE, and a similar set of parameters in the case of PYPER, for language generation.

The purpose of the **Character Model Generator** component is to generate character models to represent their personalities through their dialogue behaviors. For example, a character model can be a vector of numbers or some kind of statistical representation. This component uses the features extracted by Feature Extraction along with statistical models techniques to discover idiosyncratic aspects of a character dialogue behavior.

The **Mapping** component maps the generated character model(s) to the **Expressive NLG** component, which consists of PERSONAGE or PYPER's stylistic parameters and story domain structure. Thus the output of this portion, as well as the entire system, is composed of text versions of the dialogue.

For the **Narrative Content** component, as mentioned before, most of the current story representations bottom out at the dialogue level (Figure 1.1), and we need a better narrative representation framework that allows for manipulation of story structures at the dialogue level. We have tried using tools such as *Wide Ruled* [Skorupski et al., 2007] and *Comme il Faut* [McCoy et al., 2011], but only to discover that they do not provide a structure that represents linguistically relevant aspects of the story.

For film characters and PERSONAGE we use a scripted story from SPYFEET [Reed et al., 2011a], written for the realizer REALPRO. For TV characters and PYPER we use a story encoding tool, SCHEHERAZADE [Elson and McKeown, 2009], to help widen the number, and variety, of stories.

And finally for **evaluation**, we take generated utterances and perform user perceptual studies (qualitative evaluation) to see how well the perceived personality of generated utterances matches the modeled character. We also test model goodness (quantitative evaluation) by metrics such as the number of significant features and language model.

Our work to date suggests that generated utterances can improve author creativity, but more importantly we have discovered several areas where further research is needed in terms of narrative adaptation. First, the current state-of-the-art narrative representation contains deep representation of author goals and story structure, but such goals and intentions do not propagate down to the dialogue level. Second, significant technical work is needed to support procedural generation of character dialogue for any domain and type of character. Currently we do not know of any work on automatic generation of scene variations that can model how both content and form need to vary to reflect the radically different ways characters can interact.

Our approach of learning character models from film/TV is completely unique to our work. To our knowledge, no prior work has analyzed theatrical or film dialogue from a natural language processing perspective for the purpose of developing computational models of character [Oberlander et al., 2000, Vogel and Lynch, 2008, Pennebaker and Ireland, 2011].

This thesis is organized as follow. We describe the motivation behind the work (Chapter 2) followed by related work (Chapter 3). The **contribution of our work** is creating representative character stylistic models by extracting linguistic features related to NLG decisions, using a corpus of film (Chapter 4) and television (Chapter 5) dramatic character dialogue, to help produce believable character dialogue in an INS framework with NLG. We conclude the thesis by addressing limitations and discussing possible future work (Chapter 6).

Chapter 2

Motivation

Our focus is to extend current research on natural language generation to enable more flexible generation of interactive dialog for interactive stories. Previous research on NLG has been driven by the observation that language has a social function in addition to its use as a method for exchanging information or coordinating on tasks [Goffman, 1970, Labov, 2006, Dunbar, 1998].

Speakers use linguistic cues to project the speaker's personality, emotions, and social group, and hearers use these cues to infer properties about the speaker. While some cues appear to be produced through automatic cognitive processes, speakers may also overload their communicative intentions to try to satisfy multiple goals simultaneously, such as projecting a specific image to the hearer while communicating information and minimizing communicative effort.

We believe that Expressive Natural Language Generation (ENLG) offers the potential to address scalability issues and to produce variations in linguistic style that can manifest differences in dramatic characters. It has the potential to produce continuous stylistic variation over multiple stylistic factors by automatically learning a model of the relation between stylistic factors and properties (parameters) of generated utterances [Paiva and Evans, 2004, Paiva and Evans,

2005, Bouayad-Agha et al., 1998].

We focus on the turn variations for interactive stories, which involves:
1) developing parameters that can express the variations desired;
2) developing models that can control the parameters; and
3) developing methods to test whether the models have the desired perceptual effects.

Previous work has shown that we can control user perceptions of a character personality using rule-based models of personality traits, models that we can learn from user feedback to express combinations of personality traits [Walker et al., 1997, Mairesse and Walker, 2010, Mairesse and Walker, 2011].

This thesis aims to help make progress in some of these overall system goals by continuing our effort in creating character models through dialogue. Our work on learning models of characters from film provides evidence that parameters from our ENLG engine PERSONAGE (Table 3.2) provides many of the necessary parameters for creating a variety of models of characters. We also show experimentally that human subjects tend to perceive the generated utterances as being more similar to the character they are modeled on, than to another random character. Chapter 4 talks in more details of our work on character models from film dialogue.

Our follow-up work focuses on character dialogue from TV series, which supplies a larger set of dialogue per character that can help us better identify linguistic stylistic features. The work follows a similar pipeline to film dialogue, but with some key differences. First, we use PYPER, a Python spin-off of PERSONAGE, that provides some new controls for generation. Second, we generate character dialogue through various stories, made possible by using a narrative representation framework that allows for manipulation of story structures to the dialogue level. We use the interface created by [Rishes et al., 2013] that bridges SCHEHERAZADE,

a semantic annotation tool for stories [Elson and McKeown, 2009] and the surface realizer, RealPro, used in PyPer/Personage. And third, we use Amazon's Mechanical Turk for user perception experiment. The work is discussed in Chapter 5.

In the following chapters, we present literature review in related work (Chapter 3), character modeling with film scripts (Chapter 4) and TV episodes (Chapter 5), and finally, conclusion with possible future directions (Chapter 6).

Chapter 3

Related Work

This section describes the related research organized by components laid out in Figure 1.3: dialogue corpus, feature extraction, character model generator, narrative content, and expressive NLG.

3.1 Dialogue Corpus

Recently there has been a lot of interest in analyzing film and TV dialogue for different NLP tasks. We will first show representative datasets and research on film and TV dialogue that are related to our work, then we will look at analyses of film/TV dialogue from corpus linguistics and computational stylistics' perspective. For more information on available corpora for building data-driven dialogue system see the survey by [Serban et al., 2015a].

3.1.1 Related Corpus and Research

Table 3.1 show a list of scripted movies and TV series dialogue. The **Movie-DiC Corpus** [Banchs, 2012] is similar to ours where it contains downloaded scripts from the IMSDb website. A derived corpus, **Movie-Triples** [Serban et al.,

2015b], was created to build end-to-end dialogue systems with recurrent neural networks (RNN) and n-gram models. It contains dialogue of 3 turns between two interlocutors, which restricts the modeling data to dialogue with only two speakers. The authors also use the **SubTle Corpus**, which contains Interaction-Response pairs extracted from subtitles files. The **Cornell Movie-Dialogue Corpus** [Danescu-Niculescu-Mizil and Lee, 2011] has short conversations from movie scripts.[1] It also contains film and character metadata such as film genre and release year, and character gender and position on movie credits. The corpus was used to investigate convergence in conversational exchanges, where conversational participants tend to immediately and unconsciously adapt to each others' language styles.

For TV series, the **Corpus of American Soap Opera** [Davies, 2012b] contains transcripts of 10 American TV soap operas from 2001 to 2012.[2] The dataset, along with the Corpus of Contemporary American English and the British National Corpus, were used to compare informal, spoken English [Davies, 2012a]. The dataset does not have speaker labels. And finally, the **TVD Corpus** [Roy et al., 2014] contains transcripts from a situational comedy *The Big Bang Theory* and a fantasy drama *Game of Thrones*. It also contains crowd-sourced episode summaries, outlines, and other metadata. The dataset is used for a speaker identification task [Bredin et al., 2014]. The authors have provided scripts to reproduce the corpus locally given the user's own legal copy of the official DVD sets.[3] To the best of our knowledge, none of these corpus were used for dialogue generation to express different personalities.

[1] `http://www.cs.cornell.edu/~cristian/Cornell_Movie-Dialogs_Corpus.html`
[2] `http://corpus.byu.edu/soap/`
[3] `http://tvd.niderb.fr/corpus/`

Table 3.1: Related Movie and TV Scripted Dialogue

Corpus	# scripts	# words	Description
Movie			
Movie-DiC [Banchs, 2012]	753	6M	Movie scripts from IMSDb covering a wide range of genres.
Movie-Triples [Serban et al., 2015b]	614	13M	Derived from Movie-DiC. Dialogue of 3 turns between two interlocutors.
Cornell Movie Dialogue [Danescu-Niculescu-Mizil and Lee, 2011]	617	9M	Short conversations from film scripts, annotated with character metadata.
SubTle Corpus [Ameixa et al., 2013]	6,184	20M	Aligned interaction-response pairs from movie subtitles.
TV			
Corpus of American Soap Opera [Davies, 2012b]	22,000	100M	Transcripts of American soap operas. No speaker label.
TVD Corpus [Roy et al., 2014]	191	600k	TV scripts from a comedy (*The Big Bang Theory*) and drama (*Game of Thrones*).

3.1.2 Corpus Linguistics

There have been some work in corpus linguistics for television dialogue. One research area focuses on individual series. For example, Bednarek used *Gilmore Girls* to compare the genre dramedy to other types [Bednarek, 2011], and Quaglio compared *Friends* with unscripted conversations [Quaglio, 2009]. Another research area focuses on characterization through dialogue. For example, Bubel explored the friendship among characters in the *Sex and the City* [Bubel, 2005], and Bednarek analyzed linguistic stylistics shifts from characters from *The Big Bang Theory* [Bednarek, 2012].

While we are also interested in characterization through dialogue, our work differs from Bednarek's (and others) in that we 1) extract linguistic stylistic features based on personality studies from psychology; 2) find significant features and use them as building blocks to 3) create models using techniques such as standard scores, classification/clustering, or probabilistic/statistical approaches; and 4) apply the models to applications such as natural language generation, virtual

agents, or interactive fictions.

3.1.3 Computational Stylistics

Another related area more associated with digital humanities is computational stylistics (or stylometry), the use of quantitative methods to study writing styles to characterize authors. Principal component analysis (PCA) is often used to analyze the variations in words. It focuses on the challenge of relating features and meanings in text, which is not fixed depending on the context [Schreibman et al., 2008]. There is a local meaning in reference to the speaker, and a larger meaning in reference to the text. A popular research topic in the field is authorship attribution, which tries to generate an author profile base on his/her writings. It can be applied to many applications such as classical literary text, modern forensic text, and online reviews, just to name a few [Stamatatos, 2009]. Our work differs in that we focus on features that can be generated given our current system. With a more comprehensive system in the future, a more in-depth analysis from a stylistics perspective would possibly used to better character models.

3.2 Feature Extraction

Feature engineering is one of the key aspects in building models. Here we look at features that are relevant to characters such as emotions, language style and relationships, and group conversations.

3.2.1 Emotions through text

There have been studies on affective computing that involves the recognition of emotions from text. For example, [Neviarouskaya et al., 2009] described a

lexical rule-based approach to recognize emotions from text and an application of the developed Affect Analysis Model in *Second Life*. In [Sienkiewicz et al., 2012], analysis on the emotionally annotated dialogues extracted from IRC data demonstrate simple metrics such as that the probability of a specific emotion can be useful to predict the future evolution of the discussion. Dialogues tend to evolve in the direction of a growing entropy.

Detecting emotions through dialogue is not a main goal of this thesis, as it is often difficult to define the meaning of emotions. For example, what does it mean for a character to express anger? An obvious answer is to use content words that are generally associated with anger. However, what about the silent treatment, sarcasm, or revenge? And sometimes we might not detect the anger until a later time or through another event. Nonetheless, we see a potential in emotion identification by clustering subsets of dialogue patterns or by identifying causal/temporal relationships within conversation.

3.2.2 Language styles and relationships

Pennebaker's study of personality differences among individuals or groups of individuals is also related to our work. He developed and used a text analysis tool, Linguistic Inquiry Word Count [Tausczik and Pennebaker, 2010], to cluster words base on social, psychological, and linguistic functions. LIWC is used as part of our feature extraction, but we extract additional linguistic markers that can be used for NLG.

There are many studies that relate language styles to personalities and relationships with others. [Ireland et al., 2011] investigated where language style matching predicts outcomes for romantic relationships, using a speed dating corpus. The result appears to reflect implicit interpersonal processes central to ro-

mantic relationships. Also, the usage or pronouns as discussed in Pennebaker's book *The Secret Life of Pronouns: What Our Words Say About Us* [Pennebaker, 2011] examines how and why pronouns and other forgettable words reveal so much about us. In addition [Kacewicz et al., 2011] revealed that pronoun use reflects standings in social hierarchies. These studies are useful in providing insights on what dialogue features to extract. They may also provide insights to possible new parameters for NLG.

3.2.3 Group Conversations

It is possible to study group dynamics by looking at grouping conversations across different genres. [Pennebaker and Chung, 2012] explored how basic group processes could be revealed by people's use of pronouns, articles, prepositions, and other function words. [Pennebaker and Ireland, 2011] showed that by calculating the degree to which individuals and groups use function words across a wide variety of texts, it is possible to determine when groups are most prone to engage in violence. Again, these studies are useful in providing insights on creating new features and parameters.

3.3 Character Model Generator

Understanding, let alone creating, a model that captures a character's personality is difficult. There are many ways to build character models, and here we mention a few different approaches.

3.3.1 Archetypes

One way of character creation is to use the concept of archetype theory. It provides a number of stock characters, such as HERO, SHADOW, or CARE-GIVER, who have typical roles and personalities that can be re-used in different types of narrative. The work of [Rowe et al., 2008] produced heuristic models of character behavior using a taxonomy of 45 Master Archetypes [Schmidt, 2007], and showed how archetype models can be integrated with dialogue models for generation. Furthermore, [Munteanu et al., 2010] showed the relation between archetype and personality. It suggested that several archetypes are dominant for the ones who are in the specific developmental age, and the dominant archetype is linked with several personality types.

However, these character archetypes serve as a general guideline to help authors create characters, rather than to provide an inventory of specific parameters that define characters. For example, many protagonists in movies are clear HEROs (e.g., *Indiana Jones*, *Superman*, *James Bond*) but have radically different styles. Furthermore, characters such as *Batman*, *Zorro*, or *The Godfather* are often considered as both HERO and VILLAIN.

We believe direct modeling of characters through archetypes is more limited than our approach of learning models for specific characters from corpora. However, once individual characters are defined (in terms of their dialogue patterns), we could potentially cluster these characters into these archetypes to provide additional flexibility in generalization of character. For example, an author might be interested in dialogue spoken by all characters that can be considered as HERO or HERO-VILLAIN, as opposed to a particular individual character.

3.3.2 Latent Models

Another way to learn character types is to use latent variable models, defined as a set of mixtures over latent lexical classes, to capture the stereotypical actions of the character [Bamman et al., 2014a]. These character types, or *personas*, were built from movie plot summaries extracted from Wikipedia. They also used movies metadata drawn from Freebase [Bollacker et al., 2008]. This includes language, country, release date, genre, gender, and the actors who played them (gender, estimated age, etc.). The stereotypical actions are similar to the archetypes from [Rowe et al., 2008] such as VILLAINS. The models are learned in a manner similar of topic modeling: soft cluster words to topics (e.g., “strangle" is mostly a type of *Assault* word), soft cluster topics to personas (e.g., VILLAINS perform a lot of *Assault* actions), and hard cluster of character to personas (e.g., Darth Vader is a VILLAIN).

The authors used a slightly different model when incorporating metadata, letting the observed variables influence the character BEFORE observing their actions. An example of the clustering result in shown in Figure 3.1. The example shows *The Dark Knight*'s 4 characters (Batman, Jim Gordon, Rachel, and The Joker) being clustered in 3 different latent topics (grouped words within dashed boxes) along with “similar" characters from other movies. For example, Rachel is in the same group as Jack Dawson from *Titanic*.

The evaluations are done quantitatively by how well the same character from different movies are clustered naturally. For example, “Jason Bourne" is portrayed in three different movies, but ultimately they should belong to the same natural cluster. Variation of information and purity scores are calculated between the gold cluster and latent persona cluster.

Their follow-up work created a mixed-effect author/persona model from En-

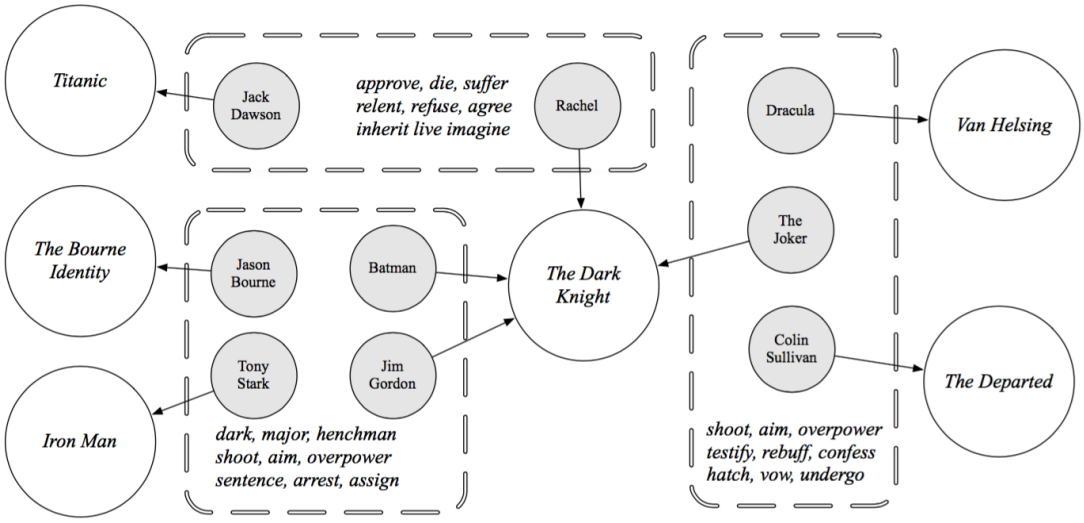


Figure 3.1: Bamman's Latent Character Model Example: *The Dark Knight*. The example shows *The Dark Knight*'s 4 characters (Batman, Jim Gordon, Rachel, and The Joker) being clustered in 3 different latent topics (grouped words within dashed boxes) along with "similar" characters from other movies. For example, Rachel is in the same group as Jack Dawson from *Titanic*.

glish novels to include author influence [Bamman et al., 2014b]. The subjective evaluation of the three models (based persona, persona regression, and mixed-effect author/persona) showed that the mixed-effect author/persona model more closely reproduce a human reader's judgments, especially when it comes to distinguishing different character types from the same author.

Learning with latent variables is a great way for persona discovery, and it is more refined than the archetypes. However the underlying method requires tuning of the numbers of latent topics and personas, which can be computationally expensive as more characters are added to the data. In addition, the paper showed that more latent topics/personas corresponds to better clustering performance. Analyzing an open range of clusters can be a daunting task.

The author stated an interesting open question: "By examining how any individual character deviates from the behavior indicative of their type, we might be

able to paint a more nuanced picture of how a character can embody a specific persona while resisting it at the same time." This is what we address in the character models, though not with topic models but by examining distinctive linguistic and stylistic features from a normalized set of features using standard scores. Our character models are more refined because they train on dialogue.

3.4 Narrative Content / Story Generation

There is an extensive amount of research in story generation (narrative content), which tends to focus on plots and character development to achieve narrative goals. Some notable story generators include TALESPIN that generates stories through inference about character goals by a carefully crafted processes [Meehan, 1977]. MINSTREL, on the other hand, is an author modeling story generator, where the actions are carried out from author's perspective [Turner, 1993, Tearse et al., 2010]. The Grail Framework offers semi-autonomous agents and a system of offering options to the player based on their previous actions within the game, but guided by authorial intent [Sullivan et al., 2010a]. It uses content-selection similar to Façde [Mateas and Stern, 2003], a fully-realized interactive drama.

Another source of creating stories comes from crowd workers writing stories in simple language [Li et al., 2012] and then learns the structure of events in a given situation, producing a script-like knowledge structure called *plot graph* [Li et al., 2013]. A second round of crowd-sourcing requested workers to write more detailed descriptions for these learned events, going into details with characters' intentions, facial expressions, and actions [Li et al., 2014]. In addition, they used the Google N-Gram Corpus and Project Gutenberg to help select different types of sentences (most/least probable, most fictional, most interesting details) and different sentiments (most positive/negative):

Example event: **Sally puts money in bag**

Most Probable	Sally put $1,000,000 in a bag.
Least Probable	Sally put the money in the bag, and collected the money from the 2 tellers next to her.
Most Fictional	Sally quickly and nervously stuffed the money into the bag.
Most Interesting	Sally quickly and nervously stuffed the money into the bag.
Most Positive	Sally continued to cooperate, putting the money into the bag as ordered.
Most Negative	Sally's hands were trembling as she put the money in the bag.

This work can be integrated into an interactive narrative system where players can perform different actions and still receive a coherence story experience [Guzdial et al., 2014]. This differs from our work in that instead of picking sentences from an existing pool of dialogue, we automatically generate/modify sentences based on distinct linguistic features extracted from a corpus, which allows for a even wider variation of dialogue.

SCHEHERAZADE is a semantic annotation and encoding tool for stories. Users construct propositions that approximate a reference text, by selecting predicates and arguments from among controlled vocabularies drawn from resources such as WordNet and VerbNet. The user then integrates the propositions into a conceptual graph that maps out the entire discourse, where the nodes represent story plots and edges represent causal and temporal relationships. Thus the entire story is encoded as linguistic representation of events (Figure 3.2). We build on and use this representation in Chapter 5.

Authoring tools such as *Wide Ruled* generate stories based on the Universe author-goal-based model of story generation [Skorupski et al., 2007]. The social AI system behind *Prom Week* [McCoy et al., 2012], *Comme il Faut (CiF)*, is a system for authoring playable social models where models of social interaction are provided for authors [McCoy et al., 2011].

Authors from [Ryan et al., 2014] tried to generate different combinations of

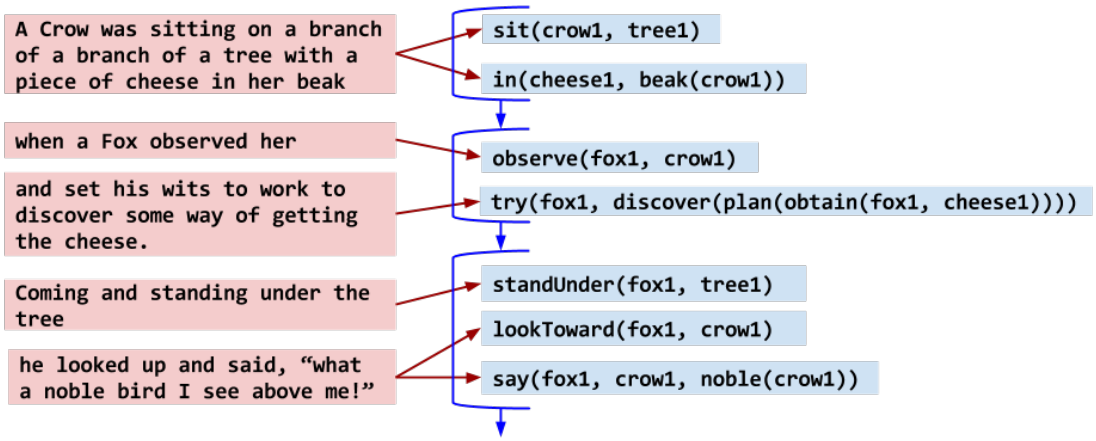


Figure 3.2: SCHEHERAZADE Encoding

Prom Week interactions by annotating its dialogue similar to the way SCHEHERAZADE does but constraints on the use to a dialogue turn. The encoded information include social-exchange identities (e.g., ask out, pick-up line), preconditions (e.g., initiator is brainy), speech acts (e.g., yes-no question, affirmative answer), and dependency lines (if any). The system then searches over the space of possible dialogue configurations given a set of constraints on the encoded information (what the exchange must do or show). Their user studies have shown that the recombinant dialogue was better at expressing game state than human-authored dialogues. Their follow-up work, currently in development, builds on the expressive-NLG system PERSONAGE [Mairesse and Walker, 2011] to create a mixed-initiative authoring tool, EXPRESSIONIST, for defining probabilistic context-free grammars that yield templated dialogue [Ryan et al., 2015].

3.5 Expressive NLG

There is an extensive amount of research in NLG, which tends to focus on linguistic structures at the sentence level. This section talks about related work in expressive NLG, including the systems we have used, PERSONAGE and PYPER

(a Python spin-off of PERSONAGE). We will first describe the traditional NLG pipeline. Note that traditional NLG's purpose is to generate informative text, while we are interested in expressive text.

3.5.1 NLG Generic Pipeline

A top-level pipeline of NLG contains three main stages: content planning, sentence planning, and surface realization (Figure 3.3). The communicative goal (as input) is transformed into a surface string of utterance. The **content planning** stage refines communicative goals, and selects and structures content. The **sentence planning** stage chooses linguistic resources (lexicon, syntax) to achieve the communication goals. The final stage, **surface realization**, uses grammar (syntax, morphology) to generate surface utterances.

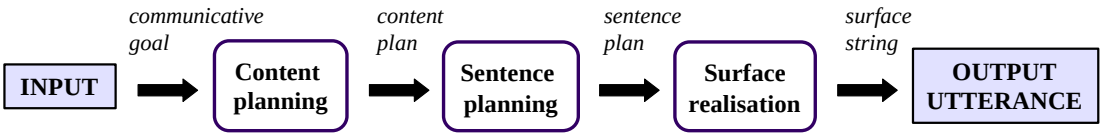


Figure 3.3: Natural Language Generation (NLG) Architecture

3.5.2 ENLG - Previous Work

ENLG has the potential to support automatic rendering of characters' linguistic behaviors, but there is still much work to be done. The primary technical aim is to integrate dynamic plot segment selection with an ENLG engine, so that (1) the player can choose to interact with any character to carry out any plot goal; and (2) the player's dialogue interaction with non-player characters (NPCs) is personalized to reflect the player's choice history.

One method is to use multi-agent story planning that involves goals that represent the outcome of a story, as opposed to the traditional plan goals representing

an agent's intended world state [Riedl and Young, 2004]. The multi-agent story planner determines agents actions to achieve a story's goal, and ensures that each agent appears to be acting intentionally. Another method involves implementing a multimodal NLG module that generates dialogues through full grammar rules, templates and canned text [Piwek, 2003]. And finally, [Callaway and Lester, 2002] presented a narrative prose generation architecture that is meant to bridge the gap between story generators and NLG systems.

Previous work on ENLG has explored parameters and models based on Brown and Levinson's theory of politeness, the Big Five theory of personality, and dramatic theories of archetypes, [Piwek, 2003, Brown and Levinson, 1987, Callaway and Lester, 2002, Mairesse and Walker, 2010, Gupta et al., 2007, Walker et al., 1997, Rowe et al., 2008, Cavazza and Charles, 2005]. While politeness and personality theories provide both character relevant parameters and models for controlling them, they do not, in any obvious way, map onto the way that authors of (interactive) stories think about character or dialogue.

Research on generating story dialogue on a turn-by-turn basis include [Cavazza and Charles, 2005] and [Rowe et al., 2008]. The first research focuses on generating story dialogue on a turn-by-turn basis is creating a dialogue generator that produces appropriate dialogue acts to guide the selection of a semantic template for the content of the utterance [Cavazza and Charles, 2005] The surface form is generated by lexicalising the semantic template to generate correct syntactic structures. While they guarantee variability of generated utterances, insertion of stylistic elements, etc., they still need to hand-craft all the linguistic resources, which can be tedious and time-consuming depending on the complexity of the scenarios.

The work of [Rowe et al., 2008] focuses on building an archetype-driven char-

acter dialogue generator that generate dialogue based on character personality and narrative history to achieve communicative goals. Their Character Dialogue Generator considers various sources of information to decompose the generation task into two separate search processes: semantic planning (search for a topic, a narrative element) and syntactic planning (search for a syntactic template to realize the topic). However, they have only focused on a few narrative scenarios with a small set of dialogues, and no empirical studies have taken place.

3.5.3 ENLG - Personage

As mentioned in the Introduction, our work is the only work to date that uses a full scale ENLG architecture for a full procedural generation of dialogue. Our ENLG engine PERSONAGE is a highly parametrizable language generator consists of 67 parameters (Table 3.2) based on the Big Five personalties that can be used to manipulate how language is generated. The sentence planning module takes an input generation dictionary consisting of a mapping from story concepts to a lexico-syntactic representation called Deep Syntactic Structures (DSyntS). DSyntS are used by RealPro, the surface realizer for PERSONAGE. Below shows an example of the DSyntS for the sentence “I would do anything for her":

```
// Sentence: I would do anything for her.
DSYNTS:
DO [mood:cond]
(    I "<PRONOUN>" [ number:sg person:1st ]
     II anything [ class:indefinite_pronoun person:3rd number:sg ]
     (    ATTR FOR []
          (    II girl [ class:common_noun gender:fem article:def pro:pro ]
)    )    )
END:
```

A primary motivation for the Big Five model is that personality trait descriptions are pervasive in descriptions of dramatic and literary character [All-

Table 3.2: PERSONAGE Generation Parameters and PYPER Support

PERSONAGE Parameter	Description	PYPER Support
Content Planning		
Verbosity	Control num of propositions in the utterance	–
Repetitions	Repeat an existing proposition	Repetition
Content Polarity	Control polarity of propositions expressed	–
Polarization	Control expressed polarity as neutral or extreme	–
Repetition Polarity	Control polarity of the restated propositions	–
Concessions	Emphasize one attribute over another	–
Concessions Polarity	Determine whether positive or negative attributes are emphasized	–
Positive Content First	Determine whether positive propositions are uttered first	–
Syntactic Template Selection		
First Person in Claim	Control the number of first person pronouns	–
Claim Polarity	Control the connotation of the claim	–
Claim Complexity	Control the syntactic complexity (syntactic embedding)	–
Aggregation Operations		
Join by Period	Two clauses joined by period	–
Relative Clause	Aggregate propositions with a relative clause	–
Conjunction	Two clauses joined with a coordinating conjunction	–
Merge with And	Merge subject and verb of two propositions	Merge
Merge with Comma	Restate proposition by repeat only the object	Merge
Object Ellipsis	Same subject; replace object of the first clause by three-dot ellipsis	–
Cue Word: With	Aggregate propositions using *with*	–
Cue Word: Also	Join two propositions using *also*	–
Cue Word: Contrast	Choices: *while, but, however, on the other hand*	–
Cue Word: Justify	Choices: *because*, *since*, *so*	–
Cue Word: Concede	Choices: *although, even if, but, though*	–
Pragmatic Markers		
Pronominalization	Replace occurrences of names by pronouns	Actor Pronouns
Negation	Negate a verb	Inverse Meaning
Exclamation	Insert an exclamation mark	emph_exclamation
In-Group	Member of the same social group	in_group_marker
Subject Implicitness	Clause needs to have the form: *NOUN has ADJ NOUN*	Subject Implicitness
Tag Question	Insert a tag question	Tag Question
Stuttering	Duplicate first letters of a name	–
Expletives	Insert a swear word	low_expletives
Near Expletives	Insert a near-swear word	near_expletives
Request Confirmation	Request confirmation	–
Initial Rejection	Begin the utterance with a mild rejection	init_reject
Competence Mitigation	Main verb subordinated to new clause	competence_ mitigation
Softeners	Soften a proposition: *kind of, sort of, somewhat, quite, rather, around, subordinate*	down_somewhat, down_quite, down_rather, down_around
Filled Pauses	Insert syntactic elements: *like, err, mmhm, I mean, you know*	all
Emphasizers	Strengthen a proposition: *really, basically, actually, just*	all
Acknowledgment	Insert an initial back-channel: *yeah, well, oh, right, ok, I see*	all
Lexical Choice		
Lexicon Freq	Average frequency of use of each content word	–
Word Length	Average num of letters of each content word	–
Verb Strength	Strength of the verbs	–

port, 1960]: *Almost all the literature of character—whether [nonfiction] or fiction, drama or biography—proceeds on the psychological assumption that each character has certain* traits *peculiar to himself which can be defined through the narrating of typical episodes from life.*

Another of Allport's observations was that traits important for describing differences in human behavior will have a corresponding lexical token, which is typically an adjective, e.g., *trustworthy, modest, friendly, spontaneous, talkative, dutiful, anxious, impulsive, vulnerable* [Allport and Odbert, 1936]. They collected 17,953 trait terms from English and identified 4,500 as stable traits. Subsequent work analyzed how traits factor together in descriptions of people, leading to a standard framework of the Big Five personality traits as a way to describe essential personality differences among humans [Norman, 1963, Goldberg, 1990].

Previous work has tested PERSONAGE both on its own and combined with text-to-speech, and facial expressions gesture engines based on the Big Five and shown that, in the case of restaurant recommendations, generated utterances are perceived by humans as expressing the intended personality traits [Neff et al., 2010, Neff et al., 2011, Bee et al., 2010, Mairesse and Walker, 2010, Mairesse and Walker, 2011]. Thus the Big Five model of personality provides a lot of parameters that can be used by any models.

3.5.4 ENLG - PyPer

PYPER is a new implementation of PERSONAGE in Python, which is currently under construction at the time of this writing [Bowden, 2016]. We gained a good understanding of the system by being one of its first users. Its hedge operations are summarized in Table 3.3. Non-hedge operations such as repetition, word change, implicitness, question/answering, polarity, and merge, are shown in Table

3.5. Most of these features are inspired by PERSONAGE, as shown in Table 3.2. The system makes it relatively easy to add additional operations as needed.

Table 3.3: PYPER Operations: Hedges

PyPer Operation	General Rule	Example *before* → *after*
Expletives		
near_expletives*		
—darn	[expletive] [Adj]	She was *able* to sing. → She was *darn able* to sing.
—oh_gosh	[expletive] [Phrase]	The cheese fell. → *Oh gosh* the cheese fell.
low_expletives*		
—damn	[expletive] [Adj]	She was *able* to sing. → She was *damn able* to sing.
—oh_god	[expletive] [Phrase]	The cheese fell. → *Oh god* the cheese fell.
Emphasizers		
emph_actually*		
—actually_start	[emph][,][Phrase]	She was able to sing. → *Actually*, she was able to sing.
—actually_end	[VerbP][,][emph]	She was able to sing. → She was able to sing*, actually*.
emph_just*		
—just_have	[NounP] [emph] [have]	*He has* gumballs. → *He just has* has gumballs.
—just_be	[NounP] [emph] [be]	The *crow is* exquisite. → The *crow is just* exquisite.
emph_typical		
–typical_end	[Phrase][.][emph][.]	The cheese fell. → The cheese fell. *Typical*.
–art_typical	[Article][emph][Noun]	The fox stood under *the tree*. → The fox stood under *the typical tree*.
emph_particularly		
–adj_particularly	[emph] [Adj]	She was *able* to sing. → She was *particularly able* to sing.
–verb_particularly	[emph][Verb]	She *was* able to sing. → She *particularly was* able to sing.
emph_technically		
–adj_technically	[emph] [Adj]	She was *able* to sing. → She was *technically able* to sing.
–adj_technically_comma	[,][emph][,][Adj]	She was *able* to sing. → She was*, technically, able* to sing.
–technically_end	[Phrase][,][emph]	She was able to sing. → She was able to sing*, technically*.
–technically_start	[emph][,][Phrase]	She was able to sing. → *Technically,* she was able to sing.
emph_literally		
–verb_literally	[emph][Verb]	She *was* able to sing. → She *literally was* able to sing.
–adj_literally	[emph][Adj]	She was *able* to sing. → She was *literally able* to sing.
–literally_end	[Phrase][.][emph]	She was able to sing. → She was able to sing. *Literally*.
emph_exclamation*	[Sentence] [!]	The cheese fell. → The cheese fell*!*
emph_you_know*	[Phrase][,][emph]	The cheese fell. → The cheese fell*, you know*.
emph_as_it_were	[Phrase][,][emph]	The cheese fell. → The cheese fell*, as it were*.
emph_basically*	[emph][,][Phrase]	The cheese fell → *Basically,* the cheese fell.
↪ same rule for: emph_essentially, emph_great		
emph_really*	[emph] [Adj]	She was *able* to sing. → She was *really able* to sing.
↪ same rule for: emph_somewhat, emph_very, emph_especially, emph_relatively, emph_largely, emph_pretty,		

* Mapped to PERSONAGE in Table 3.2.

Besides the PERSONAGE-like parameters, the system is also capable of converting monologue to dialogue, i.e., a story into conversations between two characters,

Table 3.4: PyPer Operations: Hedges 2

PyPer Operation	General Rule	Example *before* → *after*
Acknowledgement		
ack_very_well		
—very_well_start	[ack]. [Phrase]	The cheese fell. → *Very well.* The cheese fell.
—very_well_end	[Phrase].[ack].	The cheese fell. → The cheese fell. *Very well.*
ack_i_see*	[ack][,][Phrase]	The cheese fell. → *I see,* the cheese fell.
↪ same rule for: ack_well*, ack_yeah*, ack_right*, ack_oh*, ack_ok*		
Downers		
down_unfortunately		
—unfortunately	[down][,][Phrase]	The cheese fell. → *Unfortunately,* the cheese fell.
—but_unfortunately	[Phrase] [down] [Phrase]	The cheese fell *but* the fox caught it. → The cheese fell *but unfortunately* the fox caught it.
—that_unfortunately	[Phrase] [down] [Phrase]	The lady saw *that* the man saw her. → The lady saw *that unfortunately* the man saw her.
—unfortunately_end	[Phrase][,][down][.]	The cheese fell. → The cheese fell*, unfortunately.*
down_like*	[,][down][,][Adj]	She was *able* to sing. → She was*, like, able* to sing.
down_around*	[down] [%d] [Noun]	Jerry has *5 fish*. → Jerry has *around 5 fish*.
down_i_mean*	[down] [Phrase]	The cheese fell. → *I mean,* the cheese fell.
↪ same rule for: down_mmhm*, down_err*, down_i_think		
down_quite*	[down] [Adj]	She was *able* to sing. → She was *quite able* to sing.
↪ same rule for: down_somewhat*, down_rather*		

* Mapped to Personage in Table 3.2.

and other manipulations on the stories. While we use some of these features (described in Chapter 5), we leave the details to other future publications.

3.5.5 Other Methods

Recent work in statistical NLG builds a sequence-to-sequence framework by utilizing neural networks. Deep structured learning, or deep learning, attempts to model the underlying representation of data with multiple processing of many complex layers of neural networks. The mapping of the sequence-to-sequence framework has been applied to different NLP domains such as machine translation, parsing, and image captioning. Recurrent neural network has shown to be good for language modeling and dialogue in short, Twitter-like conversations. The most recent work by [Vinyals and Le, 2015] built an end-to-end, open-domain system that predicts the next sentence in a conversation. The “open domain" is trained on a large corpus of different movies’ subtitles. The chatbot seems to display some humor and attitude in its responses. More work in this area is still needed; it is

Table 3.5: PyPer Operations: Non-Hedges

PyPer Operation	General Rule	Example *before* → *after*
Markers		
In-group marker*	Choice: mate Choice: pal Choice: buddy	The cheese fell. → The cheese fell, *mate*. The cheese fell. → The cheese fell, *pal*. The cheese fell. → The cheese fell, *buddy*.
Initial rejection*	Choice: I don't know Choice: I'm not sure Choice: I might be wrong	The fox came. → *I do not know but,* the fox came. The fox came. → *I am not sure* the fox came The fox came. → *I might be wrong, but* the fox came.
Competence mitigation*	Choice: come on Choice: obviously	The cheese fell. → *Come on,* the cheese fell. The cheese fell. → *Obviously,* the cheese fell.
Soften adjective	Double negative.	She was *able* to sing. → She was *not unable* to sing.
Repetition		
Repetition*	For the other speaker.	Speaker 1: The cheese fell. → *Speaker 2: [Yeah | Right],* the cheese fell.
Word Change		
Paraphrase	Different words.	The bird *talked* to the crow about fruit. → The bird *spoke* to the crow about food.
Verbosify	Longest synonym.	The bird had cheese in its *mount*. → The bird has cheese in its *nozzle*.
Simplify	Shortest synonym.	The *airplane* flew. → The *plan* flew.
Contractfy	Group words.	The cheese fell, *did it not?* → The cheese fell, *didn't it?*
Assign speakers	Assign character name.	*Speaker 1:* The fox stood under the tree. → *Fox:* The fox stood under the tree.
Implicitness		
Actor pronouns*	Actors referred by pronouns.	*The fox* stood under the tree. → *He* stood under the tree.
Subject implicitness*	Make subject implicit.	Broken into 3 cases: That cat has a furry coat. → *The coat is furry. Her coat is furry.* *The cat's coat is furry.*
Q/A System		
Tag question*	Add tag to the end of sentence.	The cheese fell. → The cheese fell, *didn't it?*
Ask question	Convert sentence to question.	The cheese fell. → *What fell?*
Ask and answer	Asks and answers question.	The cheese fell. → *What fell? The cheese fell.*
Polarity		
Inverse meaning*	Negate verb.	The fox *ran* to the crow. → The fox *did not run* to the crow.
Merge		
Merge* **	Same subjects and verbs.	The crow has elegant talons. The crow has a pointy beak. → The crow has elegant talons *and* a pointy beak.

* Mapped to Personage in Table 3.2.
** Set manually for now.

unclear whether deep learning can be used to learn character dialogic styles, or produce entire scenes and stories.

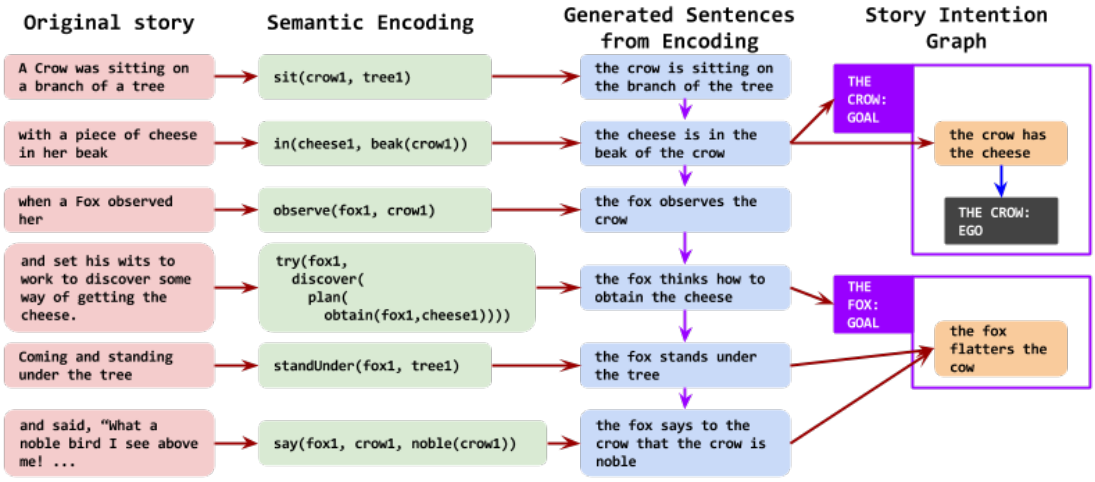
3.6 Integration of Narrative Content and Expressive NLG

There is an extensive amount of research in both story generation (narrative content) and NLG systems. Story generators tend to focus on plot and character development to achieve narrative goals, while NLG systems tend to focus on linguistic structures at the sentence level. However, there is a limited amount of work integrating the two fields. For the film dialogue experiments we use the SpyFeet storytelling system [Reed et al., 2011b], which is a custom story that will be discussed in more details in Chapter 4.5. For the TV dialogue experiments we use EST [Rishes et al., 2013], which uses Scheherazade to generate a deep representation of a story before translating it into DSyntS for further manipulations in Personage/PyPer.

ES-Translator (EST)

EST bridges the two off-the-shelf tools, Scheherazade and Personage. More specifically, it converts Scheherazade's story intention graph (SIG) to the input parameters accepted by Personage. An example of SIG for the fable *The Fox and Crow* is shown in Figure 3.4a, and an example of the translated encoding is in Figure 3.4b.

The original system produced variations of the original story told by a third person narrator. In their follow-up work [Lukin et al., 2014], an extension was built to allow direct speech to be modeled by the narrative representation, thus generating variants of stock utterances. Figure 3.5 shows a partial example of the story *The Fox and The Crow* after EST dialogic realization. The key here is to convert sentences with verbs such as *said, felt* to direct speech.



(a) Partial Story Intention Graph for *The Fox and Crow*

Scheherazade Encoding

```
sit(crow1, tree1)
```

Translated DSyntS in XML structure

```
<dsynts>
    <dsyntnode class="verb" lexeme="sit" mood="ind" rel="II" tense="past">
        <dsyntnode class="common_noun" lexeme="crow" number="sg" rel="I"/>
        <dsyntnode class="preposition" lexeme="on" rel="ATTR">
            <dsyntnode class="common_noun" lexeme="branch" number="sg" rel="II">
                <dsyntnode class="common_noun" lexeme="tree" number="sg" rel="I" />
            </dsyntnode>
        </dsyntnode>
    </dsyntnode>
</dsynts>
```

(b) Example Encoding to DSyntS

Figure 3.4: Example of SCHEHERAZADE's Story Intention Graph and Encoding to DSyntS

Single Narrator Realization	**EST Dialogic Realization**
The fox said he saw the bird.	The fox averted "I see you!"
The fox said the beauty of the bird was incomparable.	The fox alleged "your beauty is quite incomparable, okay?"
The fox said the hue of the feather of the bird was exquisite.	The fox alleged "your feather's chromaticity is damn exquisite."
The crow felt the fox flattered her.	The crow thought "the fox was so-somewhat flattering."
The fox said the crow was able to sang.	The fox said "you are somewhat able to sing, alright?"
The fox said the crow needed the wits.	The fox alleged "you need the wits!"

Figure 3.5: Example of EST Dialogic Realization with Personality: *The Fox and The Crow*

EST further adds sentence planning variations for the contingency discourse relation by manipulating the *nucleus* and *satellie* portions of the sentence (be-

Contingency Relation Example: I placed the bowl on the deck in order for Benjamin to drink the bowl's water. *–nucleus (N)*: I placed the bowl on the deck *–satellite (S)*: Benjamin (wanted) to drink the bowl's water		
Variation	**Rule**	**Output**
becauseNS	[N][because][S]	I placed the bowl on the deck *because* Benjamin wanted to drink the bowl's water.
becauseSN	[Because][S][,][N]	*Because* Benjamin wanted to drink the bowl's water, I placed the bowl on the deck.
NS	[S only][N only]	I placed the bowl on the deck. Benjamin wanted to drink the bowl's water.
N	[N only]	I placed the bowl on the deck.
soSN	[S][,so][N]	Benjamin wanted to drink the bowl?s water, *so* I placed the bowl on the deck.

Figure 3.6: Example of EST Sentence Planning Variations

causeNS, becauseSN , NS, N, and soSN). An example applied to a personal narrative from weblogs told from the first-person perspective is shown in Table 3.6. These operations allow EST to produce different tellings of stories through different voices and points-of-view, and a study has shown that such variations can manipulate the perception of characters and story engagement and interest [Lukin and Walker, 2015]. EST uses RealPro to realize the story and uses PERSONAGE (which contains RealPro) for the additional voice variations. EST-PERSONAGE and EST-PYPER provide varying narrative content that allows us to perform more interesting experiments, as we discuss in more detail in Chapter 5.

3.7 Summary: Differences of Our Approach to Previous Work

Many research work use corpora of film/TV dialogue for different NLP-related tasks. For example: speaker identification, authorship attribution, conversational participants adapt to each other's language styles, friendship characterization, scripted vs. unscripted conversations, and linguistic stylistics shifts. To the best of our knowledge, none of these corpora were used for dialogue generation to express different personalities.

Feature engineering is one of the key aspects in building models. Detecting emotions through text is a popular NLP task, but it is not a main goal of this thesis, as it is often difficult to define the meaning of emotions. There are studies that relate language styles to personalities such as using language style to predict outcome of romantic relationships. While they do not focus on language generation, they are useful in providing insights to possible new linguistic features to extract and parameters for NLG.

There are many ways to create character models. One way is through the archetype theory that uses stock characters such as HERO, SHADOW, or CARE-GIVER, so their roles and personalities that can be re-used in different stories. Another way to model character types is to use latent variable models, similar to topic modeling, by soft cluster words to topics, soft cluster topics to personas, and hard cluster of characters to personas. The personas are similar to the archetypes above such as VILLAINS. This is a great way for persona discovery and to create a more refined model than archetypes. However it requires tuning of latent topics and personas, which can be computationally expensive. We believe modeling of characters through archetypes/personas is more limited than our approach of learning models for specific characters from corpora, which is even more refined in that we train on distinctive linguistic stylistic features of the dialogue.

Narrative content/story generation tends to focus on plots and character development to achieve narrative goals. Characters' utterances are usually selected from an existing pool of dialogue. This differs from our work in that we automatically generate/modify sentences based on corpus-driven character models. Story annotation and authoring tools such as SCHEHERAZADE and *Wide Ruled* create internal representation of story but do not directly manipulate dialogue, which is a focus of our work.

Traditional NLG's purpose is to generate informative text, while we are interested in expressive text. Recent work on generating story dialogue on a turn-by-turn basis include using dialogue acts to guide the content of the utterance, creating an archetype-driven dialogue generator, or building an end-to-end system that predicts the next sentence in a conversation. However, they either involve needing to hand-craft linguistic resources or need more work in the area to determine its applicability to learning dialogic styles.

In summary, despite overlaps with these previous work, our work differs in that we:

1) extract linguistic stylistic features based on personality studies from psychology;
2) focus on features that can be generated given our current system;
3) find significant features and use them as building blocks to
4) create models using techniques such as standard scores and classification; and
5) apply the models to applications such as natural language generation.

In the following chapters we discuss our work in character modeling through film dialogue (Chapter 4) and TV dialogue (Chapter 5), and concluding the thesis with limitations and possible future directions (Chapter 6).

Chapter 4

Modeling with Film Characters

Our focus is to extend current research on natural language generation to enable more flexible generation of interactive dialog for interactive stories. Expressive Natural Language Generation (ENLG) offers the potential to produce variations in linguistic style that can manifest differences in dramatic characters. Speakers use linguistic cues to project the speaker's personality, emotions, and social group, and hearers use these cues to infer properties about the speaker. We focus on the turn variations for interactive stories, which involves:
1) developing **parameters** that can express the variations desired;
2) developing **models** that can control the parameters; and
3) developing **methods** to test whether the models have the desired perceptual effects.

This thesis aims to help make progress in some of these overall system goals by continuing our effort in creating **character models through dialogue for ENLG**. This chapter describes our work on **learning models of characters from film**. This work provides evidence that parameters from our ENLG engine PERSONAGE (Table 3.2) provides many of the necessary parameters for creating a variety of models of characters. We will also show encouraging results that human

subjects tend to perceive the generated utterances as being more similar to the character they are modeled on, than to another random character.

4.1 Introduction

We utilize our Movie Dialogue Corpus [Walker et al., 2012] to derive character models. We believe that the stylized, crafted aspects of film dialogue are actually useful for our purposes. Film dialogue is authored deliberately in order to convey the feelings, thoughts and perceptions of the character being portrayed, and the screenplay often specifies the emotion of an utterance with psychological state descriptors. In addition, the dialogue is deliberately constructed to focus the viewer's attention on the character's personality, and the key plot events involving a character and their perceptions, especially in dramatic films as opposed to action.

Here, we show how to define both character parameters and character models through an automatic corpus-based analysis of film screenplays, such as the example in Figure 4.1 from Woody Allen's *Annie Hall*. To our knowledge, no prior work has analyzed theatrical or film dialogue from a natural language processing perspective for the purpose of developing computational models of character [Oberlander et al., 2000, Vogel and Lynch, 2008, Pennebaker and Ireland, 2011].

SCENE: LOBBY of Sports Club
ALVY: Uh ... you-you wanna lift? **ANNIE:** *Turning and aiming her thumb over her shoulder* Oh, why-uh ... y-y-you gotta car? **ALVY:** No, um ... I was gonna take a cab. **ANNIE:** *Laughing* Oh, no, I have a car. **ALVY:** You have a car? *Annie smiles, hands folded in front of her* So ... *Clears his throat.* I don't understand why ... if you have a car, so then-then wh-why did you say "Do you have a car?"... like you wanted a lift?

Figure 4.1: A Scene from *Annie Hall*.

We show that we can learn at least two different kinds of models from film

dialogue. First, for individual characters we learn models that indicate significant differences in linguistic behaviors between an individual character such as Annie in *Annie Hall* and other female characters. Second, we show that we can learn models for groups of characters with classification accuracies up to 83% over a baseline of 20% based on character gender, film genre and director. While the latter method performed well, the classification models were hard to interpret and not as intuitive to drive expressive language generation. We will describe both methods in Section 4.4, but only using the individual models (the first method) to set 10 to 30 parameters of the PERSONAGE generator for human perceptual experiment in Section 4.6.

We used ontology from the movie database IMDb to define groupings of character types according to the following attributes: GENRE, DIRECTOR, YEAR, and CHARACTER GENDER. Previous work suggests that females and males in each genre might have different linguistic styles [Pennebaker and Ireland, 2011], so we use the Names Corpus, Version 1.3 (see website of Kantrowitz and Ross 1994) to label common gender names and hand-annotated the remaining characters. Note also that most films belong to multiple genres. For example, *Pulp Fiction* belongs to crime, drama, and thriller. This allows for characters to be grouped in multiple categories:

Genre	drama, thriller, crime, comedy, action, romance, adventure
Gender	male, female
Film Year	<1980, 1980<year≤1985, 1985<year≤1990, 1990<year≤1995, 1995<year≤2000, >2000
Film Director	Michael Mann, WesCraven, Steven Spielberg, Stanley Kubrick, Ridley Scott, Frank Capra, Steven Soderbergh, David Fincher, Alfred Hitchcock, Robert Zemeckis, David Lynch, James Cameron, Joel Coen, Martin Scorsese, Quentin Tarantino

The flow of our experiment is shown in Figure 4.2 and 4.3. It is summarized in

the following steps while referencing the components of our system (Figure 1.3). Each step is described in detail in the following sections of this chapter:

1. [Dialogue Corpus; Section 4.2] Collect movie scripts from IMSDb.
2. [Dialogue Corpus; Section 4.2] Parse each movie script to extract dialogic utterances, producing an output file containing utterances of exactly one character of each movie.
3. [Dialogue Corpus; Section 4.2] Select characters we wish to mimic; they must have at least 60 turns of dialogue; this is an arbitrary threshold we set to find leading roles within films.
4. [Feat Extract; Section 4.3] Extract features reflecting particular linguistic behaviors for each character.
5. [Character Model: Section 4.4] Learn models of character types based on these features.
6. [Narrative Content and Expressive NLG; Section 4.5] Use models to control parameters of the PERSONAGE engine and generate utterances.
7. [Evaluation; Section 4.6] Evaluate human perceptions of dialogic utterances generated using the character models.

4.2 Dialogue Corpus

Steps 1 to 3 consists of collecting movie scripts from IMSDb, parsing to extract dialogic utterances, and selecting "main" characters. The main characters must have at least 60 turns of dialogue; this is an arbitrary threshold we set to find leading roles within films.

Our corpus consisted of 862 film scripts from The Internet Movie Script Database (IMSDb) website[1], representing 7,400 characters, with a total of 664,000 lines of dialogue and 9,599,000 tokens. Our snapshot of IMSDb is from May 19, 2010.

[1] `http://www.imsdb.com/`

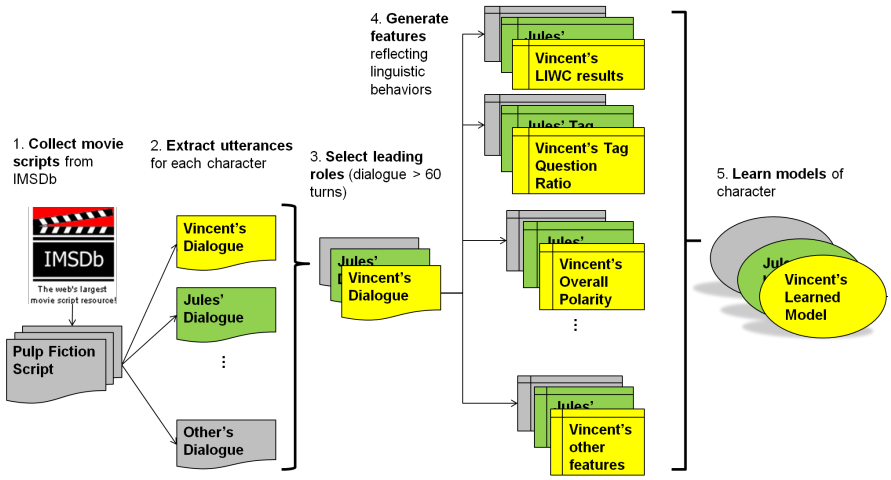


Figure 4.2: Character Creator with Film Dialogue (Step 1 to 4): Creating Features from Character Dialogue.

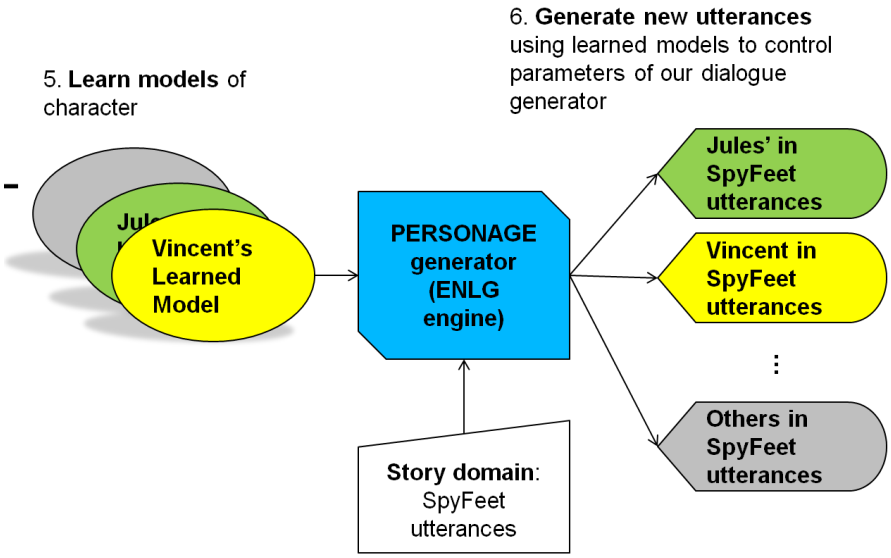


Figure 4.3: Character Creation with Film Dialogue (Step 5, 6): Creating Character Models from Features.

The content of the film corpus was released to the public through [Walker et al., 2012][2]. The original scripts contain information about the scene, the speakers, and the utterances (Figure 4.4). Each script was parsed to extract dialogic ut-

[2] `https://nlds.soe.ucsc.edu/fc`. Note that an updated version is available, created by a colleague: `https://nlds.soe.ucsc.edu/fc2`

terances, producing output files for each individual character from the film that containing only their lines. For example, *pulp-fiction-vincent.txt* contains all of the 258 lines for the character, Vincent, from *Pulp Fiction*:

VINCENT 258:

58 What so you want to know?

60 Yeah, it's legal, but is ain't a hundred percent legal. I mean you can't walk into a restaurant, roll a joint, and start puffin' away. You're only supposed to smoke in your home or certain designated places.

62 Yeah, it breaks down like this: it's legal to buy it, it's legal to own it and, if you're the proprietor of a hash bar, it's legal to sell it. It's legal to carry it, which doesn't really matter 'cause get a load of this if the cops stop you, it's illegal for this to search you. Searching you is a right that the cops in Amsterdam don't have.

... etc.

```
INT. '74 CHEVY (MOVING) - MORNING

    An old gas guzzling, dirty, white 1974 Chevy Nova BARRELS
    down a homeless-ridden street in Hollywood. In the front
    seat are two young fellas - one white, one black - both
    wearing cheap black suits with thin black ties under long
    green dusters. Their names are VINCENT VEGA (white) and JULES
    WINNFIELD (black). Jules is behind the wheel.

                JULES
        - Okay now, tell me about the hash
        bars?

                VINCENT
        What so you want to know?

                JULES
        Well, hash is legal there, right?

        VINCENT
        Yeah, it's legal, but is ain't a
        hundred percent legal. I mean you
        can't walk into a restaurant, roll a
        joint, and start puffin' away. You're
        only supposed to smoke in your home
        or certain designated places.

                JULES
        Those are hash bars?
```

Figure 4.4: A Partial Scene from the *Pulp Fiction* Unparsed, Original Script

4.3 Feature Extraction

After extracting dialogic utterances from movie scripts covering the first 3 steps, we extract features reflecting particular linguistic behaviors for each character. Procedurally generating interesting dialogue requires a large number of parameters for manipulating linguistic behavior. In step 4 of our method, in order to infer important parameters, we count features that correspond to them. Table 4.1 enumerates all feature sets, which are described in detail below.

We start by counting linguistic reflexes that have been useful in prior work characterizing individual differences in linguistic behavior due to personality and social class. While we believe that there are aspects of character not captured with this feature inventory, we attempt to quantify the extent to which they discriminate between different types of characters, and what the learned models tell us about differences in character types.

We annotated the corpus with various linguistic reflexes. A summary of these are given in Table 4.1. In some cases, we used tools that have been used previously for personality or author recognition or as useful as indicators of a person's personality, gender or social class [Mairesse et al., 2007, Furnham, 1990, Pennebaker and L.A., 1999, Pennebaker and Ireland, 2011]. We have also written new linguistic inference methods and trained a simple dialogue act tagger for the corpus.

4.3.1 Basic, Sentiment, Dialogue Act, Merge

Basic

We assumed that how much a character talks and how many words they used is a primitive aspect of character. Therefore, we counted number of tokens and turns. These, especially when considered in tandem with other features may indicate

Table 4.1: Automatically Annotated Linguistic Features for Film Dialogue

Feature Set	Description
1. Basic	Number of sentences, sentences per turn, number of verbs, number of verbs per sentence, etc.
2. Sentiment Polarity	Overall polarity, polarity of sentences, etc., using SentiWordNet[3] to calculate positive, negative, and neutral score.
3. Dialogue Act	Train Naive Bayes classifier with NPS Chat Corpus' 15 dialogue act types using simple features. We also determine "First Dialogue Act", where we look at the dialogue act of the first sentence of each turn.
4. Merge Ratio	Use regular expression to detect the merging of subject and verb of two propositions.
5. Passive Voice	Using a third party software (see text) to detect passive sentences.
6. Concession Polarity	Look for concession cues, then calculate polarity of concession portion.
7. LIWC Categories	Word categories from the Linguistic Inquiry and Word Count (LIWC) text analysis software.
8. Markers - Personage	collect words used in Personage for generation, which where selected based on psychological studies to identify pragmatic markers of personality that affect the utterance.
9. Markers - Others	Inspired by Personage words. Extended set.
10. Tag Questions	Use regular expression to capture tag questions.
11. Verb Strength	Averaged sentiment values of verbs.
12. Content Words Length	Find the average length of content words.

traits such as introversion, overall verbosity, and linguistic sophistication.

Sentiment Polarity

Positive and negative polarity were determined using SentiWordNet 3.0 [Baccianella et al., 2010]. It assigned to each synset of WordNet three sentiment scores: positivity, negativity, and objectivity. After using Stanford's Part-of-Speech Tagger, we converted Penn tags to WordNet tags. Then we approximated the sentiment value of a word with a label (no word sense disambiguation) using weights. For example, if there were three values $(v1, v2, v3)$, where $v1$ was associated with the most common sentiment value, associated with a particular word, then the score was calculated as $\frac{(1)*v1+(1/2)*v2+(1/3)*v3}{(1)+(1/2)+(1/3)}$. For more than one word

(in a sentence or entire dialogue), we simply averaged the scores. The polarity was assigned based on the range defined in Table 4.2.

Table 4.2: Polarity score with *SentiWordNet*

Polarity assigned	**Range of score** (s)
String Positive	$s \geq 2/3$
Positive	$1/3 < s < 2/3$
Weak Positive	$0 < s < 1/3$
Neutral	$s == 0$
Weak Negative	$-1/3 \leq s < 0$
Negative	$-2/3 \leq s < -1/3$
Strong Negative	$s \leq -2/3$

Dialogue Act

Different types of characters used different dialogue acts to take the initiative or in response. We trained a dialogue act tagger on the NPS Chat Corpus 1.0 [Forsyth and Martell, 2007], and apply it to each turn's utterances. The 15 dialogue act types are shown in Table 4.3 with examples. A related feature is to look at the dialogue act of the first sentence of each turn only.

Table 4.3: NPS Chat Corpus Dialogue Act Examples

Dialogue Act	**Example**
Accept	*yeah it does, they all do*
Bye	*night ya'all.*
Clarify	*i meant to write the word may...*
Continuer	*and thought I'd share*
Emotion	*lol*
Emphasis	*Ok I'm gonna put it up ONE MORE TIME*
Greet	*hiya hug*
No Answer	*no I had a roommate who did though*
Other	*0*
Reject	*u r not on meds*
Statement	*Yay...democrats have taken the house!*
System	*JOIN*
Wh-Question	*why do you feel that way?*
Yes Answer	*why yes I do lol*
Yes/No Question	*cant we all just get along*

Merge Ratio

To detect merging of sentences (merge of subject and verb of two propositions), we used regular expression to capture various patterns such as: *verb + noun + conjunction + noun.*

Passive Voice

Passive sentences were detected using a third party script.[4] These scripts implemented the rule that if a *to-be* verb is followed by a non-gerund, the sentence is probably in passive voice.

Concession Polarity

Here we separate an utterance by a concession word/phrase cue. The two separated parts are the *main* portion (usually to the left of the cue) and the *concession* portion (usually follows the cue). We then calculate the polarity score for the concession portion. Some example cues include: *although, however, whereas, on the other hand, all the same, notwithstanding, nonetheless, nevertheless, despite*, etc.

For example the sentence, *the two rivals were nevertheless united by the freemasonry of the acting profession*, can be broken down as:
– cue: *nevertheless*
– main portion: *the two rivals were*
– concession portion: *united by the freemasonry of the acting profession*

LIWC Categories

The Linguistic Inquiry Word Count (LIWC) [Pennebaker et al., 2001] tool provides a lexical hierarchy that tells us how frequently characters use different types of

[4] `http://code.google.com/p/narorumo` → `source/browse/trunk/passive`

words such as words associated with anger or happiness, as well as more subtle linguistic cues like the frequent use of certain pronouns. Examples of the LIWC word categories are given in Table 4.4.

Table 4.4: Sample LIWC Word Categories and Examples

Category	Example	Category	Example
Anger words	*hate, kill, pissed*	Metaphysical	*God, heaven, coffin*
Physical state	*ache, breast, sleep*	Inclusive words	*with, and, include*
Social processes	*talk, us, friend*	Family members	*mom, brother, cousin*
Past tense verbs	*walked, were, had*	Ref. to friends	*pal, buddy, coworker*
Causation	*because, know, ought*	Discrepancy	*should, would, could*

4.3.2 Pragmatic Markers

Pragmatic markers are important parts of linguistic style [Brown and Levinson, 1987] so we developed ways to count them. These include both the categories of pragmatic markers and individual word count/ratio. These markers are inspired by the ones used in PERSONAGE. The categories and examples are shown in Table 4.5.

4.3.3 Tag Questions, Content Words, and N-Grams

A **tag question** turns a statement sentence into a question and attaches it the end of the original sentence. For example, the sentence “This is a book" becomes the tag “isn’t it", and the whole sentence becomes “This is a book, isn’t it?" To find tag questions we used regular expressions to parse sentences. **Verb strength** was determined by averaging the sentiment scores (via SENTIWORDNET) of all verbs. To find the **average content word length**, we first used WORDNET’s tag to find content words (noun, adjective, adverb, and verb), and then averaged the length of words (number of letters).

Table 4.5: Different Word Categories and Examples

Category	Examples
taboo	*fuck, shit, hell, damn, darn, goddamn, ass, bitch, sucks, piss, cunt, cocksucker, motherfucker, tits, jesus christ, oh my god*
near swear	*hell, heck, darn, sucks, jesus christ, oh my god*
sequence	*first, second, third, next, last, finally, then, during, now, as, "while", already, recent, earlier, later, until, by, following, soon, at the same time*
opinion	*think, feel, believe, should, must, seem, good, better, best, wonderful, nice, beautiful, bad, worse, worst, terrible, more, most, less, least, greatest*
aggregation	*with, also, because, since, so, although, but, though, even if"*
soften	*somewhat, quite, around, rather, i think, it seems, it seems to me*
emphasizers	*really, basically, actually, just*
acknowledge	*yeah, right, ok, oh, i see, oh well*
fillers	*i mean, you know*
concession	*but, yet, although, though, while, whereas, however, notwithstanding, nonetheless, nevertheless, despite, even though, on the other hand, all the same, all the same time, even if, in spite of*
concede	*although, but, though, even if*
justify	*because, since, so*
contrast	*while, but, however, on the other hand*
conjunction	*for, and, nor, but, or, yet, so*
in-group	*pal, mate, buddy*
relative	*who, whom, whose, that, which, when, where, why*
misc	*", with", ", also", kind of, sort of*

4.4 Character Model Generator

Step 5 of our experiment flow involves learning models of character types based on the linguistic stylistic features extracted from film characters' dialogue. We initially explored whether it is possible to classify characters by different character groupings (gender, film genre, etc.). While we got good accuracies in identifying groups of characters this way, the classification models were not useful for generation because they often do not contain enough features to express personalities through speech patterns via drive PERSONAGE. For example, to distinguish between movies by Tarantino and Hitchcock one only needs to look for the presence (or absence) of swearing words.

We then explored another method where we focused on significant differences between a character and a group of characters. While there are many different

ways we could learn such models, here we estimate models using the Z-score, a statistical measurement of a score's relationship to the mean in a group of scores, to find significant features. The two methods can be summarized as follow:

1. For **groups of characters** we learn different **classification models** with accuracies up to 83% over a baseline of 20% based on character gender, film genre and director. These models turned out to be NOT useful for generation. They will not be discussed beyond this section.

2. For **individual characters** we learn **Z-score models** that indicate significant differences in linguistic behaviors between an individual character such as Annie (*Annie Hall*) and other female characters. These models turned out to be useful for generation, which will be described later in Section 4.5 and 4.6.

4.4.1 Character Models from Classification Models

We trained the classification models for groups of characters using the *ZeroR* method (majority class; used as baseline) and J48 pruned decision tree (an implementation of the C4.5 algorithm [Quinlan, 2014]), with 10-fold cross validation, through the data mining software Weka [Hall et al., 2009]. We first selected a subset of relevant features using the search method "best first, forward". The feature subset evaluator used was CFS (correlation-based feature subset). We reported results for average classification accuracy over all folds:

Top Classification Results for Character Styles Learned Using J48 Decision Trees

Group: Categories	Selected	Test Case	Size	Baseline	J48 Accuracy
Genre: drama, thriller, crime, comedy, action, romance, adventure	**Genre, Gender**	Drama Female vs. Adventure Male	813	50.43%	74.05%
		Family Male vs. Biography Male	181	49.72%	74.03%
		Western Male vs. Animation Male	78	48.72%	71.79%
Directors: Mann, Craven, Spielberg, Kubrick, Scott, Capra, Soderbergh, Fincher, Hitchcock, Zemeckis, Lynch, Cameron, Coen, Scorsese, Tarantino	**Five Directors**	Mann vs. Hitchcock vs. Lynch vs. Cameron vs. Tarantino	108	18.35%	64.22%
		Mann vs. Lynch vs. Hitchcock vs. Kubrick vs. Zemeckis	103	19.42%	53.40%
	Gender, Director	Male: Mann, Capra, Fincher, Cameron, Tarantino	87	22.99%	66.67%
		Female: Scott, Capra, Fincher, Cameron, Coen	34	29.40%	50.00%
Film Period: now–2005, 2005–2000, 2000–1995, 1995-1990, 1990–1985, 1985–1980, before 1980	**Gender, Years**	Male: now–2005, 2005–2000, 2000–1995, 1995–1990, before 1980	4041	20.29%	83.37%
		Female: now–2005, 2005–2000, 2000–1995, 1995–1990, before 1980	1134	20.28%	76.37%

The above results show that, for the two-class **Genre X Gender** categories of character, we can distinguish the two using a binary classification model with accuracies $> 70\%$ compared to baselines of $\approx 50\%$. These learned models focused on particularly salient stylistic differences. For example, Western males can be distinguished from Animation males by 1) the use of shorter words; 2) the use of causal process words; and 3) less use of the phrase “I think".

The remaining results involving five-way discriminatory models for combinations of five directors, gender and years were much more complex. The accuracies are good given the baseline $\approx 20\%$. We could easily develop distinct character models for different directors and gender/director combinations. Also interestingly, the results showed that the year of the film had a large impact on style, and that combinations of gender and time period could be discriminated with

accuracies as high as 83%.

Even though the classification method gave us high accuracies on distinguishing various groups of film characters, the results were not useful for NLG. For example, to distinguish between movies by Tarantino and Hitchcock one only needs to look for the presence (or absence) of swearing words. For now we will not consider the classification method further.

4.4.2 Character Models from Z-Scores.

The Z-score, or the standard score, indicates the number of standard deviations away from the mean. We first create vectors of features representing individual characters, and then derive distinctive features for that character by normalizing the feature counts against a representative population. For example we can normalize Annie (*Annie Hall*) against all female characters. For each feature x_i, the normalized value z_i is calculated as:

$$\frac{x_i - \overline{x_i}}{\sigma_{xi}} \tag{4.1}$$

Any Z-score greater than 1 or less than -1 is more than one standard deviation away from the mean. Z-scores greater and less than ± 1.96 are statistically significant differences of that character compared to other characters. The normal distribution and corresponding standard deviations / Z-scores are shown in Figure 4.5.

There is a choice about the population of characters used for the normalization, i.e. which set of characters are used to calculate the mean $\overline{x_i}$ and the standard deviation σ_{xi}. For example, for a female character, obvious choices include all the characters, all the female characters, or all the female action characters. Here we normalize individual characters against all of the characters of the same gender.

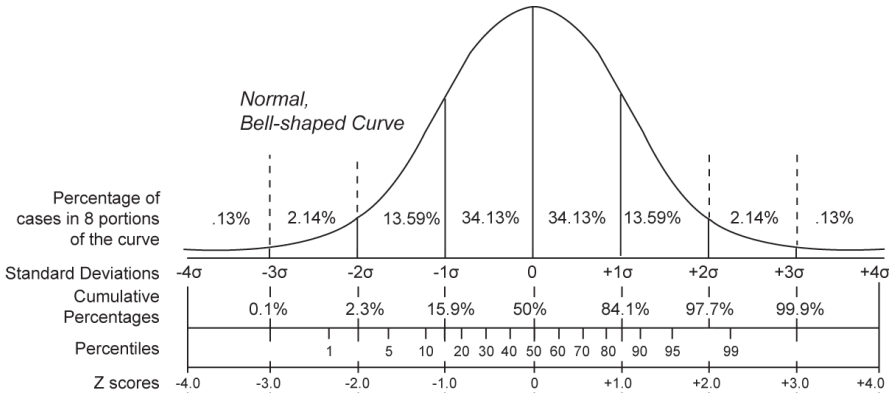


Figure 4.5: Normal Distribution and Corresponding Standard Deviations, Z-Scores Scales

Any z-score greater than 1 or less than -1 is more than one standard deviation away from the mean. Z-scores greater and less than +/-1.96 indicate significant differences of the use of that linguistic feature by that character compared to other characters. However for experimental purposes we map any z-score greater than 1 or less than -1 into one or more PERSONAGE generation parameters.

We build character models by comparing individual characters to a population of same gender characters and extracting attributes with Z-scores >1 or <-1, i.e., more than one standard deviation away from the mean. These high and low Z-scores indicate the unique attributes that made particular characters stood out from his/her gender population. A small set of characters with selected features and examples is shown in Table 4.6. The full set of Z-scores is shown in the Appendix Table .1.

For example, the model for Annie (*Annie Hall*) showed that she used many nonfluencies (LIWC-Nonfl) such as *Um* and *Uh*. She said *yes, yeah* a lot (LIWC-Assent, *yeah* ratio). She used a lot of tag questions. In addition, she utilized pragmatic marker transformations for emphasis and hedging such as *really*, *sort of*, and *I think*. A specific male character, Col. Landa (*Inglourious Basterds*), used words like *oh well* and *however* more frequently than male characters in general.

Table 4.6: A Small Set of Characters with Selected Features and Examples

Char	Director (Film)	Z-score >1	Examples
Annie (Female)	Woody Allen (*Annie Hall*)	LIWC-Nonfl (10.6)	*mhm; u-uh; huh; uh; er*
		LIWC-Assent (4.8)	*okay; yeah*
		tag Q ratio (3.3)	*I know, it's pretty silly,* ***isn't it?***
		really ratio (1.6)	***Really****, do you think so,* ***really****? You* ***really*** *thought it was good?*
		sort of ratio (1.4)	*I don't do it very often, you know, just* ***sort of****, er ... relaxes me at first. Well, I do commercials,* ***sort of*** *...*
		yeah ratio (1.2)	***yeah****,* ***yeah****!;* ***yeah****, you know something? Oh,* ***yeah****?*
		LIWC-I (1.2)	*Well, it ruins it for* ***me*** *if you have grass. If it has my name on it, then I guess it's* ***mine****.* ***I*** *think* ***I*** *know exactly what you mean.*
Lisa (Female)	A. Hitchcock (*Rear Window*)	*because* ratio (3.3)	*Only* ***because*** *it's expected of her.*
		it seems ratio (1.9)	*Not quite -* ***it seems****.*
		even if ratio (1.7)	***Even if*** *I had to pay, it would be worth it - just for the occasion.*
		I mean ratio (1.6)	***I mean*** *when he looked at the note?*
		LIWC-Discrep (1.4)	*According to you, people* ***should*** *be born, live and die on the same –. What* ***would*** *you think of starting off with dinner at the "21"? I* ***could*** *get you a dozen assignments tomorrow.*
		kind of ratio (1.4)	*You can't buy that* ***kind of*** *publicity.*
		LIWC-Incl (1.3)	*I'm sending up a plain, flat silver one -* ***with*** *just your initials engraved. For yourself -* ***and*** *me.*
		right ratio (1.2)	*I'm all* ***right****.;* ***Right*** *off the paris plane. Going on* ***right*** *here; At least you can't say the dinner isn't* ***right****.*
		just ratio (1.2)	*I'm* ***just*** *warming some brandy. I'm* ***just*** *that women don't leave jewelry behind when they go on a trip.*
Landa (Male)	Q. Tarantino (*I. Basterds*)	*oh well* ratio (9.0)	***Oh well****, must not of been important.*
		however ratio (5.0)	***However****, all I have to do, is pick up that phone right there.*
		quite ratio (3.3)	*However, I've been lead to believe you speak English* ***quite*** *well?*
		actually ratio (3.2)	***Actually*** *why he would hate the name, "The Hangman", is baffling to me.* ***Actually*** *quite the contrary, it will be met with reward.*
Mitch (Male)	A. Hitchcock (*The Birds*)	*it seems* ratio (18.6)	***It seems*** *to be a pattern, doesn't it?*
		right ratio (1.4)	*That's* ***right****.* ***Right*** *next to the one you already had.*
		LIWC-Family (1.4)	*These are for my* ***sister****. Her* ***father*** *owns a big newspaper in San Francisco.*
		I think ratio (1.1)	***I think*** *I can handle Melanie Daniels by myself.*

He also was less likely to hedge, to assent, to talk about himself, than other male characters.

There are three different ways in which our mappings of feature counts to parameters could be incomplete: (1) We have a parameter in PERSONAGE but none of the features we count are good indicators that we should use it (e.g.,

Competence Mitigation); (2) We have a feature we count but no parameter to map it to (e.g., LIWC-Discrep); or (3) There is some aspect of linguistic style that is essential to expressing a particular character's style, but we currently do not have a feature that indicates when a character has that style, nor do we have any existing parameter that could manifest that linguistic reflex.

4.5 Generate Utterances from Character Models

Step 6 of our experiment flow uses learned character models to control parameters of our dialogue generator. There are two parts to this section: 1) the narrative content, SpyFeet, for which we intend to produce character dialogue for, and 2) the mapping of character models to Personage parameters.

4.5.1 Narrative Content - SpyFeet

For film dialogue, we tested our approach in the context of *SpyFeet* [Reed et al., 2011a, Reed et al., 2011b], SpyFeet is a playable prototype of a storytelling system designed to increase agency and reduce authorial burden. The two main components are 1) story representation that adapts to player choices and 2) a dynamic dialogue generation that allows different revealings of a single story event.

An example dialogue for two characters, Sparrow and Otter, is shown in Table 4.7, and the story is summarized in Figure 4.6. DSynts were created with the utterances of *SpyFeet* characters. The idea was to apply character models learned from film dialogue and apply them to a completely different story domain. Then we performed evaluation to see if the generated utterances were perceived to have a similar style to the original characters.

Table 4.7: Sample Dialogue Creation for Two SpyFeet Characters

Character	Sparrow	Otter
Hand-authored dialogue sample	Hello there! Hello? Hey! Hi! I can see that you're new to this. Look down. The small brown bird? That's me.	Gosh, I don't think anyone knows more about your Aunt Elsebeth than Sparrow?
General traits NLG parameters	gregarious, social, impulsive, flighty Repetition, exclamation, short sentences	playful, child-like, eager, curious Expletives, in-group address terms, tag questions, disfluencies
NLG sample	Oh I mean, you must thwart Cartmill. You need to stop Cartmill. No one is worse than Cartmill.	Well, mmhm... no one is worse than Cartmill, so Cartmill cannot be permitted to continue.

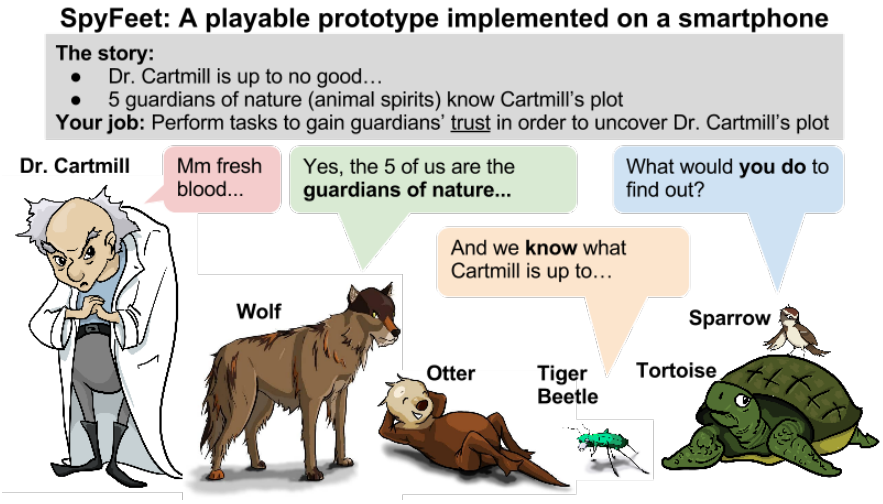


Figure 4.6: SpyFeet Story

4.5.2 Mapping to Expressive NLG - Personage

In the Expressive NLG module we use character models learned from film to drive parameters of Personage to generate character-like utterances in a different story domain. Personage was originally built for the restaurant recommendation domain but was expanded to support generic dialogue for the SpyFeet story. Our current heuristic is to map each feature, or combinations of features, to these generation parameters. For example, *Pragmatic Markers* features in Table 4.1 each corresponded to aggregation parameters in Personage or pragmatic transformations, such as insertion of emphasizers or hedges.

So far we tested our character models learned from Z-scores in the context

of SpyFeet to control parameters of the Personage generator. This is done by mapping our character models' attributes to Personage parameters with different values. We started with a default character model that represents "neutral" personality. As each character model only had a **subset of all possible attributes that are significant**, these attributes modify their corresponding Personage parameters.

For example, the *Annie Hall* characters, Alvy and Anny, had significant z-scores (2.12 and 3.28 respectively) for the tag question ratio feature. The tag question ratio represented the placement of phrases like *you know?* and *would you be?* at the end of sentences. The feature value maps to a value of 1.0 for the Personage tag question insertion parameter, causing utterances generated using the Annie or Alvy character models to include the use of tag questions. The Annie and Alvy models also lead to significant z-scores for the LIWC-WC feature (word count), which maps to the verbosity parameter in Personage. **The significant z-score value for LIWC-WC caused an increase in the verbosity parameter for the Alvy and Annie models**, and as a result, utterances generated using these models had more words than those from models with lower verbosity values such as Vincent or Indy.

Each attribute of the character model could be mapped to one or more Personage parameters, and vice versa. Currently character models are mapped to Personage parameters using a **weighted average of features**. For example, Annie's (*Annie Hall*) notable non-fluencies (LIWC-Nonfl) was used to control Personage parameters *Filled Pauses* and *Stuttering*. Table 4.8 shows a partial mapping of her other features. Sample full character models are shown in Table 4.9. Each model parameter in the left-hand side of table was described in Table 3.2.

Table 4.8: Partial Map of Learned Character Model for Annie (*Annie Hall*) to PERSONAGE Parameters: Weighted Average of Features.

PERSONAGE parameter	**Description**	**Sample mapped features from character model**	**Annie**
Verbosity	Control # of propositions in the utterances	Number of sentences per turn, words per sentence	0.78
Content polarity	Control polarity of propositions expressed	Polarity-overall, LIWC-Posemo, LIWC-Negemo, LIWC-Negate	0.77
Polarization	Control expressed polarity as neutral or extreme	1 if polarity-overall is strong negative or positive	0.72
Concessions	Emphasize one attribute over another	Category-concession	0.83
Positive content first	Determine whether positive propositions – including the claim – are uttered first	Accept-ratio, Accept-first-ratio	1.00

The result of applying these models of characters to SPYFEET utterances is illustrated in Table 4.10. It shows different variations of the same set of utterances through character models of Alvy and Annie (*Annie Hall*), Indy (*Indiana Jones*), and Vincent (*Pulp Fiction*). The differences across are shown in different colors. Alvy's stuttering shows off in words such as *st-strange* and *fr-fr-friends*. Annie's tag questions comes across with *is he?* Indy's character is relatively uneventful, and Vincent's swearing and eccentric personality is expressed through *damn*, *oh god*, and *everybody knows that*.

4.6 Evaluation

So far we have discussed (1) learning models of character linguistic style from film dialogue screen plays and (2) using the learned models to control the parameters of PERSONAGE, an expressive language generation engine. The last component of our experiment flow is evaluating our learned character models. We describe two methods of evaluation. First, we look at model "goodness" base on "number of parameters learned". Second, an experiment on human perceptions of

Table 4.9: Sample Learned Character Models

Parameter	Alvy	Annie	Indy	Marion	Mia	Vincent
Content Planning						
Verbosity	.79	.78	.36	.65	.49	.18
Repetitions	.38	0	0	0	.28	.51
Content Polarity	.09	.77	.15	.15	.15	.50
Polarization	.39	.72	.22	. 21	.22	.57
Repetitions Polarity	.54	.79	.29	.29	.29	.64
Concessions	.83	.83	.83	.89	.89	.58
Concessions Polarity	.56	.26	.56	.26	.26	.49
Positive Content First	0	1.00	0	0	0	1.00
Initial Rejection	0	0	0	0	0	0
Syntactic Template Selection						
Use of First Person in Claim	.39	.6	.39	.39	.39	.54
Claim Polarity	.57	.57	.57	.49	.56	.50
Claim Complexity	.71	.31	.47	.15	.56	.56
Aggregation Operations						
Period	.05	.04	.24	.04	.24	0
Relative Clause	0	0	.95	.97	.53	.3
With cue word	.44	.51	.05	.34	.31	.25
Conjunction	.30	.21	.22	.18	.08	0
Merge	.61	.87	.83	.65	.59	.77
Also cue Word	.12	.05	.05	.05	.07	.05
Contrast-Cue word	.76	.85	0	.84	.76	.96
Justify-Cue Word	.97	.48	0	.61	.61	.45
Concede-Cue Word	1.00	0	0	1.00	0	.25
Merge With Comma	.27	.42	.5	.5	.32	.5
Pragmatic Markers						
Stuttering	.54	.54	.04	.04	.54	.09
Pronominalization	1.00	1.00	1.00	.75	.5	1.00
Negation	0	0	0	0	0	0
Softener Hedges	1.00	1.00	0	1.00	0	0
Emphasizer hedges	0	1. 0	0	0	1.00	0
Acknowledgements	1.00	1.00	0	0	1.00	0
Filled Pauses	1.00	1.00	0	0	0	0
Exclamation	0	0	0	1.00	0	1.00
Expletives	0	0	0	0	0	1.00
Near Expletives	0	0	0	0	0	0
Tag Question	1.00	1.00	0	0	1.00	0
In-Group Marker	0	0	0	1.00	0	0
Lexical Choice						
Lexicon Frequency	.19	.19	.28	.19	.55	.18
Lexicon Word Length	.21	.13	.21	.13	.78	.28
Verb Strength	.59	.59	.5	.61	.5	.49

the character utterances created using these models. We test our approach in the context of a prototype role playing game SpyFeet [Reed et al., 2011a, Reed et al., 2011b], a game intended to support dynamic quest selection and dialogue generation, determined by user choices and user relationships with game characters [Sullivan et al., 2010b].

Table 4.10: Utterances for SpyFeet generated using Film Character Models

Film, Character, and Generated Utterances			
***Annie Hall*: Alvy**	***Annie Hall*: Annie**	***Indiana Jones*: Indy**	***Pulp Fiction*: Vincent**
- I don't know. People say Cartmill is st-strange, alright? Err... on the other hand, I don't rush to judgment.	- Come on, I don't know, do you? People say Cartmill is strange while I don't rush to um.. judgment.	- I don't rush to judgment, but people say Cartmill is strange.	- Basically, I don't rush to judgment. On the other hand, people say Cartmill is strange, he is strange.
- Right, I am not sure, would you be? I will tell something you because you br-brought me cabbage.	- I don't know. I think that you brought me cabbage, so I will tell something to you , alright?	- I will tell something you since you brought me cabbage.	- Yeah, I can answer since you brought me cabbage that.
- Oh I am not sure. Wolf wears a hard shell. On the other hand, he is ge-ge-gentle, isn't he?	- Yeah, I am not sure, would you be? Wolf wears a hard shell but he is really gentle.	- Wolf is gentle but he wears a hard shell.	- Everybody knows that Wolf wears a hard shell. He, however, is gentle.
- I see, I don't know. I respect Wolf, wouldn't you? He, however, isn't my close friend.	- I see, I am not sure. Obviously, I respect Wolf. However, he isn't my close friend, is he?	- Wolf isn't my close friend. But I respect him.	- I respect Wolf. However, he isn't my damn close friend.
- Yeah, I don't know. Sparrow conveys excitement to my life, so I am fr-fr-friends with her.	- Come on, I am not sure. Because Sparrow brings excitement to my life, I am friends with her, you see?	- I am friends with Sparrow since she brings excitement to my life.	- Oh God I am friends with Sparrow because she brings excitement to my life.

4.6.1 Quantitative Method: Model Goodness Metric

Our goal here is to be able to compare different types of models of character. We hope to develop a series of quantitative metrics that will be predictive of the quality of a model, obviating the need to do detailed perceptual experiments on each model. However at the moment, our primary quantitative metric for model quality is the number of parameters in the model that indicate significant differences in linguistic style, altogether, and at each level of statistical significance.

It would be expected that the more dialogue turns are used for the model, the better the model would be. This suggests that combining character dialogue utterances in some way could be useful: it would give us a large corpus of utterances for character type as opposed to tokens. Thus using our current quantitative evaluation method we examined the effect of number of dialogue turns on the number of significant attributes in the models.

We look at male characters Ace (*Casino*) with 747 turns, Indiana Jones (*Indiana* series) with 776 turns, and Jack from *Fight Club* with 626 turns. For female characters, we chose Carrie from *Sex and the City* with 518 turns, Hermione from the *Harry Potter* series with 481 turns, and Jackie from *Stepmom* with 444 turns. They were chosen based on the large number of their dialogue turns compared to other characters. The turns were randomized and separated into incrementing segments of roughly 100 turns. For example, Indiana has 776 turns, separated into 7 segments where segment 1 contains the first 110 turns, segment 2 contains the first 110 turns plus the next 111 turns, etc. Segment 7 contains all 776 turns. See Table 4.11 as an example.

Table 4.11: Number of Significant Attributes based on Dialogue Turns for Z>3 or Z<-3

Segment (turns)	Male Characters			Female Characters		
	Ace (Casino)	Indy (Indiana Jones)	Jack (Fight Club)	Carrie (Sex & the City)	Hermione (Harry Potter)	Jackie (Stepmom)
1(∼100)	15	13	15	12	18	15
2(∼200)	16	17	15	14	20	16
3(∼300)	17	16	16	14	21	17
4(∼400)	18	17	16	15	21	18
5(∼500)	18	19	17	16	–	–
6(∼600)	18	19	16	–	–	–
7(∼700)	18	20	–	–	–	–

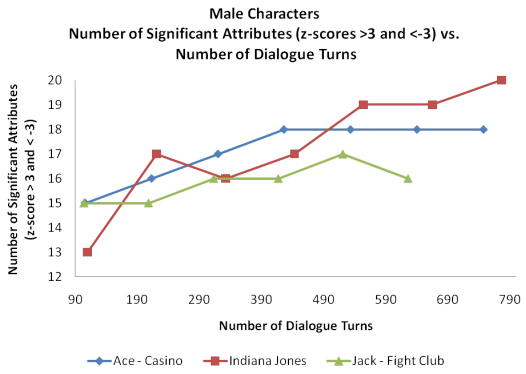


Figure 4.7: Trends for Male Characters

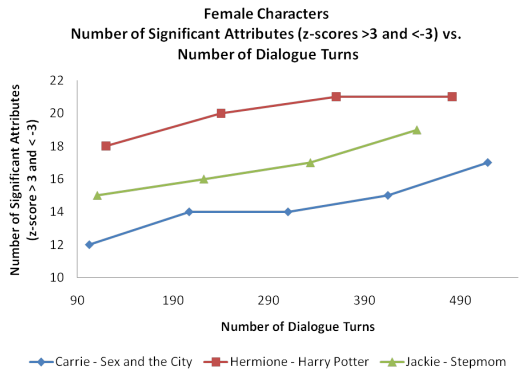


Figure 4.8: Trends for Female Characters

We found that the number of significant attributes $z > 1$ and $z < -1$ stayed

relatively the same regardless of the number of turns for all characters. However, as we increase the cutoff to $z > 3$ and $z < -3$, there is a trend that the more turns a character has, the more significant attributes there are in the resulting learned model (Table 4.11). Figure. 4.7 and 4.8 shows trends for male and female characters. From these individual characters' plots, we can see that $z > 2$ and $z < -2$, as well as $z > 3$ and $z < -3$, show an upward trend.

4.6.2 Qualitative Evaluation: User Perceptual Experiment

We test our automatically creating "character voices" based on a corpus-based statistical expressive language generation engine that is trained on the IMSDb corpus of film screen plays [Lin and Walker, 2011b]. These automatically created character voices are also intended to reveal subtext about character personality and emotion. Our results, described here, demonstrate that an approach of this sort can produce significant and recognizable variations in linguistic style, even using corpora as small as the utterances of a single character in a screenplay.

We believe that the corpus-based approach is a much stronger first step than, for example, asking authors to directly tune the parameters of a natural language generation engine. The expertise required to understand the parameters involved, and their interactions, is far removed from the expertise of creative writing — while authors are quite accustomed to presenting character voices through examples, or describing a character's voice as similar to another's (or a blend of familiar voices). Further, being able to explore a landscape of utterances produced through examples could also prove a powerful tool for novice (or even expert) authors who are considering possibilities for character voices.

In this study, we showed that 1) our expressive generation engine can operate on content from the story structures of the role-playing game SpyFeet; 2)

PERSONAGE parameter models can be learned from film dialogue; and 3) the parameter models learned from film dialogue are generally perceived as being similar to the modeled character.

4.6.3 Experimental Setup

Our goal was to test the character models and mappings as described in previous sections of this chapter. The simplest way to do this was to ask human participants to rate a set of utterances produced using different models in terms of similarity of linguistic style to the mimicked character.

Using the PERSONAGE generator, we generate dialogic utterances for the characters in the story of the SPYFEET RPG, using 6 film character model (3 male and 3 female). Using all the utterances generated, for each film character model, we generate a page showing the participant (1) three scenes from each of the original films, illustrating the utterance styles of the character; and (2) **all** generated utterances using all of the film character models. Then we ask participants to judge on a scale of 1 to 7 how **similar** the generated utterance is to the style of the film character as illustrated in the three scenes. Participants are instructed to use the whole scale, and thus effectively **rank** the generated utterances for similarity to the film character. An example for the character Marion (*Indiana Jones*) is shown in Figure 4.9.

We examine the correlations between perceptions of the film character's original utterances (Annie, Alvy, Vincent, Mia, Indiana, Marion) and SPYFEET utterances that were generated using the learned models of the film character. Thus we are interested in testing the hypothesis: *Utterances generated using character models will be perceived as being more similar to that character than utterances generated using another randomly selected character model.*

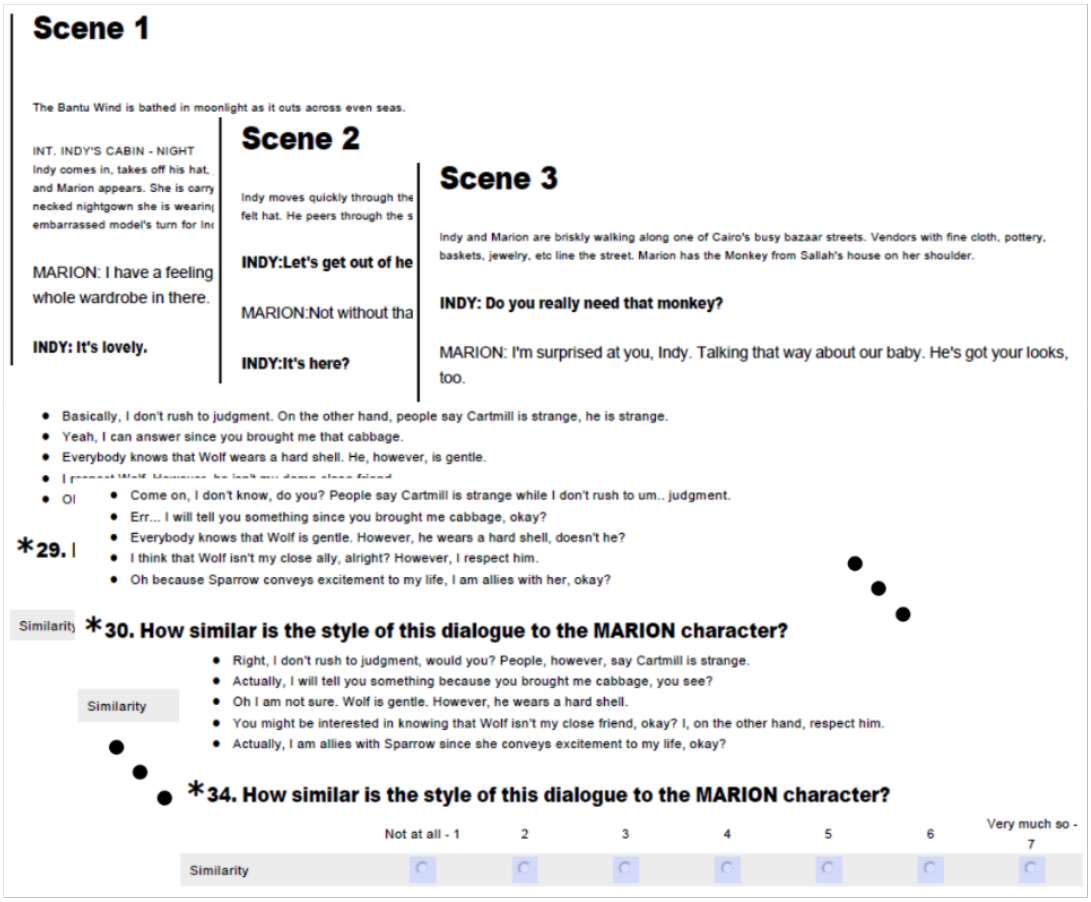


Figure 4.9: Perceptual Experiment Example
There are 3 scenes from original films illustrating utterance styles of Marion (*Indiana Jones*) and 6 sets of generated utterances in the story of SpyFeet from 6 character models; user rates each set in terms of its similarity to the Marion character.

4.6.4 Experimental Results

The experiment consisted of 29 subjects (13 female and 16 male, ages ranging from 22 to 44) who participated in a web-based experiment. Our prediction was that utterances generated using character models would be more similar to that character than utterances generated using another randomly selected character model. Table 4.12 shows the average similarity score judgments between utterances produced with a particular character model and the utterances of that

character in the original film.

Table 4.12: Average Similarity Scores between Character and Character Models.

Highest similarity scores between the character and character models are shown in **blue/bold**. Significant **differences** ($p<0.05$) between the character and character models are shown with *.

Character (Film)		Character Model					
		Alvy	Annie	Indy	Marion	Mia	Vincent
1. Alvy	(*Annie Hall*)	**5.2**	4.2 *	2.1 *	2.6 *	2.8 *	2.3 *
2. Annie	(*Annie Hall*)	4.2	**4.3**	2.8 *	3.4 *	3.9	2.9 *
3. Indy	(*Indiana Jones*)	1.4 *	2.2 *	**4.5**	2.8 *	3.3 *	3.8 *
4. Marion	(*Indiana Jones*)	1.6 *	2.8	3.7	3.1	4.1 *	**4.2** *
5. Mia	(*Pulp Fiction*)	1.7 *	2.4 *	**4.3**	3.2	3.6	**4.3**
6. Vincent	(*Pulp Fiction*)	2.1 *	3.2 *	4.5	3.5 *	3.6 *	**4.6**

For example, row 1 of Table 4.12 shows the judgments for the similarity of utterances generated with each character model to the utterances of the Alvy character in the original *Annie Hall* screen play. Similarity scores are scalar values from 1...7. The strongest possible result would be a matrix with 7's along the diagonal and 0's in all the other cells, i.e., only utterances generated with a particular character's model would be judged as being at all similar to that character. **In general, what we are looking for is a matrix with the highest values along the diagonal.**

We conducted paired t-tests comparing the similarity scores of each other character model to the similarity scores for the matching model (e.g., we compared similarity scores for utterances generated using Alvy's model to utterances generated using Indy's model, collected in the context of the participant looking at the screenplay for *Indiana Jones*.).

The results in Table 4.12 show that 4 out of 6 character models were able to generate utterances, through the SpyFeet story, and still being perceived as being similar to the original film character. The 4 characters are Alvy and Annie from *Annie Hall*, Indy from *Indiana Jones*, and Vincent from *Pulp Fiction*. Their

similarity scores are the strongest between the character and its corresponding character model.

We further analyze the results in terms of significant differences (p-value$<$0.05) produced by the paired t-test. For the two *Annie Hall* characters, we see that utterances generated using the **Alvy** model (row 1) are significantly more similar to Alvy (average similarity score = 5.2) than utterances generated using any other character model. Utterances generated using the **Annie** model (row 2) are significantly more similar to Annie (average similarity score = 4.3) than utterances generated with the models for Indy/Marion/Vincent ($p<0.05$) but not different than utterances generated with the models for Alvy/Mia ($p\geq0.05$).

For the two *Indiana Jones* characters, utterances generated using the Indy model (row 3) are significantly more similar to Indy (similarity score = 4.5) than utterances generated using any other character model. Utterances generated using the Marion model (row 4), unfortunately, seem to be confused with the ones generated by other character models. The utterances are significantly more similar to Marion than utterances generated using Alvy/Mia/Vincent models ($p<0.05$), but not different than the Annie/Indy model ($p\geq0.05$).

For the two *Pulp Fiction* characters, utterances generated using the Mia model (row 5), unfortunately, also seem to be confused with the ones generated by other character models. The utterances are significantly more similar to Mia than utterances generated from the Alvy/Annie models, but not different than those using models for Indy/Marion/Vincent. The fact that the model for the Mia character was trained on the fewest number of utterances (she has only 81 lines in the film) could contribute to the lack of perceivable differences. Last but not least, utterances generated using the Vincent model (row 6) are significantly more similar to Vincent than utterances generated using Alvy/Annie/Marion/Mia models, but

not different than the Indy model.

4.7 Summary

We have shown how we learn character models from film dialogue in order to define character types for SpyFeet. The models are based on features that can be extracted fully automatically from screenplays. The learned models identify features, and the corresponding generation parameters in Personage that can be used to produce utterances in dialogue whose style should match a particular character or group of characters. The summary of flow is shown in Figure 4.10. It may also be possible to generate even better models to better represent these characters, by creating additional relevant features and gathering additional film scripts.

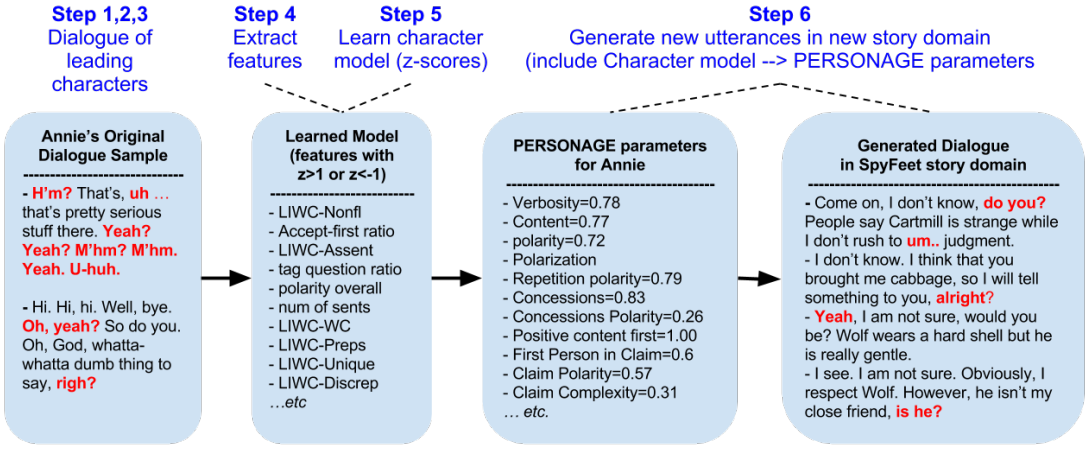


Figure 4.10: Summary of Character Modeling with Film Dialogue

If deeply interactive stories are to feature dialogue, we must move beyond a model of pure hand authoring. As stories vary in terms of the events that take place, the characters that are present, the dynamic states of relationships between characters, and so on, we must be able to dynamically generate dialogue that

reflects and drives the state of the fictional world while expressing character in a manner controllable by an author. But asking authors to, for example, specify the parameter settings for a complex natural language generation engine is at odds with the skill sets and approaches of most authors, whether experts or beginners.

In this chapter we have demonstrated the first step toward an alternative approach: developing models of character linguistic style from examples, specifically using character utterances in film scripts. Our results are encouraging, showing that utterances generated in a different domain (that of an role-playing game) recognizably display important subtext for character personality as well as style that is more similar to the modeled character than to others (though, perhaps unsurprisingly, characters from the same genre or film are often more similar to each other than to others). We believe that the current results are encouraging and suggest further exploration of both the effects of corpus size and ways to group characters as ways to generate better models.

Our follow-up work in the next chapter focuses on characters with a relatively fixed set of personalities and a larger set of dialogue: **television characters**. Having a larger set of dialogue helps us better identify linguistic stylistic features, and it allows us to explore social dynamics (speaker-addressee) in terms of their speech patterns. It follow a similar pipeline to film characters but with some key differences. First, it uses PyPer, a Python spin-off of Personage, that provides some new controls for generation. Second, instead of using one story framework like SpyFeet, many stories are used in character dialogue generation. And finally, it uses workers from Amazon's Mechanical Turk for the perceptual experiment.

Chapter 5

Modeling with Television Characters

Recall that our focus is to extend current research on natural language generation to enable more flexible generation of interactive dialog for interactive stories. Expressive Natural Language Generation (ENLG) offers the potential to produce variations in linguistic style that can manifest differences in dramatic characters. Speakers use linguistic cues to project the speaker's personality, emotions, and social group, and hearers use these cues to infer properties about the speaker. We focus on the turn variations for interactive stories, which involves:

1) developing **parameters** that can express the variations desired;

2) developing **models** that can control the parameters; and

3) developing **methods** to test whether the models have the desired perceptual effects.

This thesis aims to help make progress in some of these overall system goals by continuing our effort in creating **character models through dialogue for ENLG**. This chapter describes our follow-up work to previous chapter's learning character models from film dialogue (Chapter 4): **learning models of charac-**

ters from TV series. More specifically we pick *The Big Bang Theory* due to its mainstream popularity and author's familiarity with it.

While we follow a similar pipeline as film dialogue, creating character models and performing qualitative/quantitative evaluation, this work differs in that:
1) the extracted features are updated;
2) we use PyPer, a Python spin-off of Personage, that provides new controls for generation;
3) the narrative content is much more varied (different stories), made possible by the EST interface [Rishes et al., 2013] that bridges Scheherazade's story representation to RealPro, the surface realizer used by PyPer;
4) language model is used in the quantitative evaluation; and
5) Amazon's Mechanical Turk (MTurk) is used for user perception experiment.

This work provides evidence that using EST and a new ENLG PyPer provide at least as much, if not better, control of many of the necessary parameters for creating a variety of models of characters. The results are, again, encouraging in that human subjects tend to perceive the generated utterances as being more similar to the character they are modeled on, than to another random character.

5.1 Introduction

TV series provides another rich resource of hand-crafted dialogue. It is perhaps closer to our daily conversation than the movie dialogue due to the differences in the two mediums. Stories in movies must be told within a fixed time duration, while each TV episode tells only part of the whole story. Visuals and sounds are important aspects of movies in order to create an exciting experience for the audience, and therefore a major component of the story is told through non-dialogue. TV, on the other hand, provides more room for dialogue, character

development, and story telling through longer periods of time.

Social psychologists such as [Berne, 1996] tells us that dialogue in "real life" is often scripted based on our culture. Fiske's analysis of television culture stated that the reality is already encoded by the codes of our culture [Fiske, 2002]. He defined "codes of television", where the code is a rule-governed system of signs, whose rules and conventions are shared amongst members of a culture, and which is used to generate and circulate meanings in and for that culture. These codes are linked between producers, texts, and audiences, to produce a sense of reality.

Through television programs, characters and their dialogue give a reflection of the reality, a peek into its culture. Characters' words provide us a glimpse to their thoughts, their hearts, and their personalities. We fall in love with their souls and learn to care for them. "On the most mundane level, dialogue helps us distinguish one person from another... But the more significant use of dialogue is to make characters substantial, to hint at their inner life... Dialogue lines are explicitly designed to reveal character" [Kozloff, 2000].

Conversations in the media serve as a representation of ordinary conversational activities. In particular we focus here on conversations in situational comedies (sitcoms). They showcase different social relations of a particular (sub)culture through fun and relatable characters. For example, *Friends* (NBC, 1994-2004) and *Sex and the City* (HBO, 1998-2004) aired around the same time and both describe a group of friends living in New York City dealing with career and relationship issues. However, *Friends* had mixed company in their 20's and addressed a variety of topics, while *Sex* had all women in their 30's and 40's and focused mostly around sex. In general, a successful show balances between the uniqueness of the new material to stand out from the crowd and the amount of stereotypes to relate to different audience groups.

Even though our work focuses only on speech patterns, we believe it will help future work in exploring group dynamics. The remaining sections of this chapter follow the experimental flow of film dialogue. Below shows the steps for creating character models from film and TV as a comparison. The corresponding chapter sections are also noted:

	Creating character models from ...		
Step	**Film**	**TV**	Section
1	Collect movie scripts (from IMSDb)	Collect TV scripts (fan transcription)	5.2
2	Extract utterances for each character	← Same	5.2
3	Select leading roles (dialogue>60 turns)	← Same: leading roles are obvious	5.2
4	Generate features reflecting linguistic behaviors	← Same: some updated features	5.3
5	Learn models of character using Z-score	← Same	5.4
6	Generate new utterances using learned models to control parameters of our dialogue generator story: SpyFeet generator: Personage	← Same story: Scheherazade and EST generator: PyPer	5.5

5.2 Dialogue Corpus

We look for fan-transcribed *The Big Bang Theory* (BBT) scripts online and parse the HTML files to get the scenes, speakers, and utterances, just like what we did for film scripts. An example of a scene is shown in Figure 5.1. We collected seasons 1-4 and partial season 5.

Scene: A corridor at a sperm bank.

Sheldon: So if a photon is directed through a plane with two slits in it and either slit is observed it will not go through both slits. If it's unobserved it will, however, if it's observed after it's left the plane but before it hits its target, it will not have gone through both slits.
Leonard: Agreed, what's your point?
Sheldon: There's no point, I just think it's a good idea for a tee-shirt.
Leonard: Excuse me?
Receptionist: Hang on. *[Working on a cross-word puzzle]*
Leonard: One across is Aegean, eight down is Nabakov, twenty-six across is MCM, fourteen down is move your finger phylum, which makes fourteen across Port-au-Prince. See, Papa Doc's capital idea, that's Port-au-Prince. Haiti.

Figure 5.1: One Scene from the TV Series *The Big Bang Theory*

The show centers around 5 characters, 4 of them (all male) are scientists/engineers working at Caltech in Pasadena, California, and 1 (Penny) is a waitress. The comedy's theme focuses on the contrast between the geekiness of the male characters and Penny's social skills. Two additional female characters, both scientists, were introduced as love interests to two main male characters, and have since became main characters themselves. A summarized description of each character[1] is shown in Table 5.1.

Table 5.1: Summarized Character Description of *The Big Bang theory*

Character	**Role**	**Other info & personality**
Leonard	PhD; experimental physicist at Caltech	roommate with Sheldon; nerd who loves video games, comic books, and D&D; eventually has a relationship with Penny.
Sheldon	PhD; theoretical physicist at Caltech	roommate with Leonard; poor grasp of others' feelings; boasts his superior intelligence, belittle others, appears childlike; into rituals and routines, compulsion to complete things, dislike physical contact; eventually has a relationship with Amy.
Penny	Waitress at The Cheesecake Factory	neighbor of Sheldon/Leonard; friendly and outgoing, has common sense and social awareness; untidy, drinks alcohol; close friends with Bernadette and Amy; eventually has a relationship with Leonard.
Howard	MEng; aerospace engineer at Caltech	fancies himself as a ladies' man; many pickup lines with little success; overly confident about women; close to Raj; eventually has a relationship with Bernadette
Raj	PhD; astrophysicist at Caltech	come from a wealthy family in New Delhi; has trouble talking to women unless he drinks alcohol (or thinks that he drank alcohol); innocent; has a feminine tastes; close to Howard.
Bernadette	PhD microbiology	originally a waitress with Penny; later work as a pharmaceutical rep; sweet and good-natured but short-fused; can lash out when provoked; can be competitive; eventually has a relationship with Howard.
Amy	PhD neurobiology	met Sheldon through online dating site; shares Sheldon-like qualities, but became more social after befriending Penny and Bernadette; eventually has a relationship with Sheldon.

[1] `https://en.wikipedia.org/wiki/The_Big_Bang_Theory`

5.3 Extracted Features

Similar to features extracted for film dialogue (Table 4.1), here we list the old and new ones used for TV dialogue in Table 5.2. Note that some descriptions are repeated here to make the chapter more self-contained.

Table 5.2: Automatically Annotated Linguistic Features for TV Dialogue

Feature Set	Description
1. Basic	**[Film only]** Number of sentences, sentences per turn, number of verbs, number of verbs per sentence, etc. **[TV only]** Tokens per sentence, tokens per utterance, etc., plus words from different types of emotion and other psychological categories from the Nodebox English Linguistics library.
2. Sentiment Polarity	**[Film and TV]** Overall polarity, polarity of sentences, etc., using SentiWordNet[2] to calculate positive, negative, and neutral score.
3. Dialogue Act	**[Film and TV]** Train Naive Bayes classifier with NPS Chat Corpus' 15 dialogue act types using simple features. We also determine "First Dialogue Act", where we look at the dialogue act of the first sentence of each turn.
4. Merge Ratio	**[Film and TV]** Use regular expression to detect the merging of subject and verb of two propositions.
5. Passive Voice	**[Film and TV]** Using a third party software (see text) to detect passive sentences.
6. Concession Polarity	**[Film and TV]** Look for concession cues, then calculate polarity of concession portion.
7. LIWC Categories	**[Film and TV]** Word categories from the Linguistic Inquiry and Word Count (LIWC) text analysis software.
8. Markers - Personage	**[Film and TV]** collect words used in Personage for generation, which where selected based on psychological studies to identify pragmatic markers of personality that affect the utterance.
9. Tag Questions	**[Film and TV]** Use regular expression to capture tag questions.
10. Verb Strength	**[Film and TV]** Averaged sentiment values of verbs.
11. Content Words Length	**[Film and TV]** Find the average length of content words.
12. Markers - Others	**[TV only]** Inspired by Personage words. Extended set.
13. Hedges	**[TV only]** Collect words from a list of pre-defined hedges and their categories. LACKOFF hedges.
14. Repeating Verbs	**[TV only]** Find verbs that are repeated used in a turn.
15. BIGRAMs	**[TV only]** Top 10 bigrams.
16. Part-of-Speech BIGRAMs	**[TV only]** Top 10 POS bigrams.

In the **basic** set of features, we assumed that how much a character talks and how many words they used is a primitive aspect of character. Therefore, we counted number of tokens and turns. These, especially when considered in tandem with other features may indicate traits such as introversion, overall verbosity, and linguistic sophistication.

The NodeBox English Linguistics library[3] categorize words as emotional, persuasive, or connective. It uses Ogden's basic English words (express 90% of concepts in English) and the Regressive Imagery Dictionary, which assigns scores to primary, secondary, and emotional process thoughts in a text:

Primary: free-form associative thinking involved in dreams and fantasy

Secondary: logical, reality-based and focused on problem solving

Emotions: expressions of fear, sadness, hate, affection, etc.

The library can also categorize words as emotional, persuasive or connective. For example, the `is_basic_emotion()` command returns `True` if the given word (e.g., cheerful) expresses a basic emotion (anger, disgust, fear, joy, sadness, surprise). The `is_persuasive()` command returns `True` if the given word is a "magic" word (you, money, save, new, results, health, easy, etc.). The `is_connective()` command returns `True` if the word is a connective (nevertheless, whatever, secondly, etc.; and words like *I, the, own, him* which have little semantical value).

The library also allows us to check the emotional value of a word. For example, the command `noun.is_emotion()` guesses whether the given noun (e.g., anger) expresses an emotion by checking if there are synonyms of the word that are basic emotions.

Positive and negative sentiment polarity were determined using SentiWordNet 3.0. It assigned to each synset of WordNet three sentiment scores: pos-

[3] `https://www.nodebox.net/code/index.php/Linguistics#rid`

itivity, negativity, and objectivity. After using Stanford's Part-of-Speech Tagger, we converted Penn tags to WordNet tags. Then we approximated the sentiment value of a word with a label (no word sense disambiguation) using weights. For example, if there were three values $(v1, v2, v3)$, where $v1$ was associated with the most common sentiment value, associated with a particular word, then the score was calculated as $\frac{(1)*v1+(1/2)*v2+(1/3)*v3}{(1)+(1/2)+(1/3)}$. For more than one word (in a sentence or entire dialogue), we simply averaged the scores. The polarity was assigned based on the range defined in Table 4.2.

Different types of characters used different **dialogue acts** to take the initiative or in response. We trained a dialogue act tagger on the NPS Chap Corpus 1.0, and apply it to each turn's utterances. The 15 dialogue act types are shown in Table 4.3 with examples. A related feature is to look at the dialogue act of the first sentence of each turn only.

To detect **merging of sentences** (merge of subject and verb of two propositions), we used regular expression to capture various patterns such as: *verb + noun + conjunction + noun*. **Passive sentences** were detected using a third party script.[4] These scripts implemented the rule that if a *to-be* verb is followed by a non-gerund, the sentence is probably in passive voice.

To find the **concession polarity** of a sentence, we separate an utterance by a concession word/phrase cue. The two separated parts are the *main* portion (usually to the left of the cue) and the *concession* portion (usually follows the cue). We then calculate the polarity score for the concession portion. Some example cues include: *although, however, whereas, on the other hand, all the same, notwithstanding, nonetheless, nevertheless, despite*, etc.

For example the sentence, *the two rivals were nevertheless united by the freemasonry of the acting profession*, can be broken down as:

[4] `http://code.google.com/p/narorumo` → `source/browse/trunk/passive`

– cue: *nevertheless*

– main portion: *the two rivals were*

– concession portion: *united by the freemasonry of the acting profession*

The **LIWC** tool provides a lexical hierarchy that tells us how frequently characters use different types of words such as words associated with anger or happiness, as well as more subtle linguistic cues like the frequent use of certain pronouns. Examples of the LIWC word categories are given in Table 4.4.

Pragmatic markers are important parts of linguistic style [Brown and Levinson, 1987] so we developed ways to count them. These include both the categories of pragmatic markers and individual word count/ratio. These markers are inspired by the ones used in PERSONAGE. The categories and examples are shown in Table 4.5.

A **tag question** turns a statement sentence into a question and attaches it the end of the original sentence. For example, the sentence "This is a book" becomes the tag "isn't it", and the whole sentence becomes "This is a book, isn't it?" To find tag questions we used regular expressions to parse sentences.

Verb strength was determined by averaging the sentiment scores (via SENTIWORDNET) of all verbs. To find the **average content word length**, we first used WORDNET's tag to find content words (noun, adjective, adverb, and verb), and then averaged the length of words (number of letters). To find **repeating verbs** we parse and part-of-speech tag the words in a turn, and then look at the verbs to see if it is used more than once.

Last but not least we find **bigrams** and **part-of-speech bigrams** of utterances. Given a sequence of text, n-gram is a consecutive sequence of n items (words, letters, etc.). In our case we look at bigrams with words as items. So a sentence "This is a book" contains bigrams of (this is), (is a), and (a book). Part-

of-speech bigrams is similar to regular bigrams, except here we look at bigrams of words' part-of-speech.

5.4 Character Model Generator

Similar to character models from film (Chapter 4.4), we calculate the standard score (z-value) for each feature to build character models from BBT. Recall that the standard score is $z = (x - \mu)/\sigma$, where μ and σ are the mean and standard deviation of the population. The population depends on the experiment (e.g., all characters or all conversations). Significant features have $|z| \geq 1$, representing at least one standard deviations away from the mean. Character models are composed of these significant features.

The population in our case is all 7 main characters: Leonard, Sheldon, Penny, Howard, Raj, Bernadette, and Amy. We extracted features with $|z| \geq 1$ (one standard deviation away from the mean). In the actual mapping of features to PyPer we focused on the positive values only. Number and examples of significant features for each characters are shown in Table 5.3. We see that for $|z| \geq 1$, Sheldon, Penny, Bernadette and Amy have over 200 significant features. Sheldon, more specifically, has close to 400 significant features. When we narrow down the standard deviation to $z \geq 2$, significant features for Bernadette and Amy decreased by over 85%, Leonard, Penny, Howard, and Raj decreased by 70%, and Sheldon decreased by 54%. In addition, Sheldon (180) and Penny (68) have the most significant features, indicating their polar opposites personalities.

Table 5.3: Number and Examples of Significant Features for *The Big Bang Theory* Characters

Speaker	$\|z\| \geq 1$	$\|z\| \geq 2$	**Example Features for $z \geq 1$ (i.e., positive z-values only)**
Leonard	172	54	words:[*even if, nevertheless, whereas, even though*], Dialogue Act–{Greet, Bye}, LIWC-{Causation, Impersonal Pronouns}, hedges per sentence, connect words, concept words
Sheldon	394	180	words: [*all the same, although, despite, however, nevertheless, on the other hand, whereas, more or less, though, all, yet*], passive-ratio, important words per utt/sent, LIWC–{Inhibition, Prepositions, Number, Quantifiers}
Penny	232	68	words:[*nevertheless, even if, while, even though, on the other hand, yet*], connect words, emotional words, Dialogue Act–{Greet, Bye}, swear/near swear words, LIWC–{Adverbs, Present Tense, Dictionary Words}
Howard	133	41	words:[*although, even if, whereas*], LIWC–{Hear, See, Third Person Singular}, concept words, in-group words, hedges-per-sent
Raj	179	51	words:[*on the other hand, however, despite, though, also, even though, but*], in-group words, LIWC–{Conjunctions, Third Person Plural, See}, hedges per sentence
Bernadette	283	43	persuasive words, emotional words, conceptual words, words:[*even though, yet, while*], Dialogue Act–emphasis, LIWC–{Personal Pronouns, Second person, Auxiliary Verbs, Function Words, Past Tense}
Amy	246	43	LIWC–{Quantifiers, UniqueWords, FutureTense, Causation}, RID Emotion words, Dialogue Act - Continuer, opinion words, words:[*though, but*]

5.5 Generate Utterances from Character Models

This section corresponds to step 6 of the experiment flow, which uses learned character models to control parameters of our dialogue generator. Instead of having two separate components (narrative content, ENLG) like the film dialogue, we use the framework ES-Translator (EST) (described in related work in Chapter 3) that integrates the two by bridging the two off-the-shelf tools, Scheherazade and Personage. More specifically, EST interfaces between Scheherazade and the RealPro surface realizer. Since PyPer uses RealPro as well, we are able to

use EST with PyPer.

The overall workflow for this portion is to 1) annotate stories using Scheherazade; 2) use EST to automatically translate annotated stories to DSyntS; 3) PyPer reads and manipulates DSyntS to add expressive elements, and 4) send "expressive" DSyntS to RealPro for generation.

Our narrative content comes from 7 different stories, 1 fable and 6 blog stories from [Gordon et al., 2007]. Only the first one, the Fox and the Crow fable, is described in this section. Remaining stories are in the Appendix:

1. The Fox and the Crow Fable: Figure 5.2
2. The Garden Blog Story: Figure .1 (Appendix)
3. The Protest Blog Story: Figure .2 (Appendix)
4. The Squirrel Blog Story: Figure .3 (Appendix)
5. The Bug Blog Story: Figure .4 (Appendix)
6. The Employer Blog Story: Figure .5 (Appendix)
7. The Storm Blog Story: Figure .6 (Appendix)

The Fox and the Crow fable is shown in the top portion of Figure 5.2, labeled "Original". Some phrases are bolded to show how they were annotated and translated through Scheherazade and EST to the results shown in the left portion. We can see that many complicated sentences have been broken down into shorter ones. Note that some additional descriptions (adjectives) were added in order to provide enough search space for PyPer to exercise enough expressive parameters, so that characters' personalities will come through in different variations of the story.

For example, the sentence *a Crow was sitting on a branch of a tree with a piece of cheese in her beak* was broken down into 6 smaller sentences: *The crow sat on*

the tall tree's branch. The crow has elegant talons. The crow has a good beak. The crow has ugly wings. The crow has bad eyes. The delicious cheese was in the crow's beak. A few of them were added to increase expressivity and variability for PyPer.

One unique functionality of PyPer is its ability to convert monologue to dialogue (M2D). Even though it is still under development, we see the results in the middle portion, where the fable content is divided evenly between two speakers. The results is that the two speakers seem to be having a conversation about a fox and a crow. For example, there are a total of 26 sentences from the annotation. Assuming each dialogue turn is composed of 3 or 4 sentences, this gives us about 8 turns total, 4 per speaker. The middle portion of Figure 5.2 show s the 8 turns of the fable as dialogue.

And finally, after adding stylistic features, which may change the number of overall sentences and therefore turns, the expressive version of the dialogue is shown in the right portion. The yellow highlight shows the expressive snippets added to the dialogue. For example, one of the stylistic parameter is to convert a statement to a question. So the statement *the delicious cheese was in the crow's beak*) becomes the question *what was in her beak?*

The Fox and the Crow (Original)

A Crow was sitting on a branch of a tree with a piece of cheese in her beak when a Fox observed her and set his wits to work to discover some way of getting the cheese. **Coming and standing under the tree** he looked up and said, "What a noble bird I see above me! Her beauty is without equal, the hue of her plumage exquisite. If only her voice is as sweet as her looks are fair, she ought without doubt to be Queen of the Birds." **The Crow was hugely flattered by this**, and just to show the Fox that she could sing she gave a loud caw. **Down came the cheese**, of course, and the Fox, snatching it up, said, "You have a voice, madam, I see: what you want is wits."

Scheherazade & EST	**PyPer: Monologue to Dialogue (M2D)**	**PyPer: M2D + Stylistic Parameters**
The crow sat on the tall tree's branch. The crow has elegant talons. The crow has a good beak. The crow has ugly wings. The crow has bad eyes. The delicious cheese was in the crow's beak. The hungry fox observed the crow. The fox has good claws. The fox has bad teeth. The fox tried to discover for the fox to get the cheese. **The fox came. The fox stood under the tall tree.** The fox looked toward the crow. The fox said the fox saw the crow. The fox said the crow's beauty was incomparable. The fox said the crow's plumage's hue was exquisite. The fox said if the crow's voice's sweetness was the crow's appearance's fairness the crow would undoubtedly be the every noble bird's queen. **The crow felt for the fox to flatter the crow.** The gullible crow loudly cawed. The crow wanted to show the fox for the crow to be able to sing. **The coveted cheese fell.** The cunning fox snatched the cheese. The cunning fox snatched the crackers. The fox is full now. The fox said the crow was able to sing. The fox said the crow needed the wits.	Speaker 1: **The crow sat on the tall tree's branch.** The crow has elegant talons, and a good beak. The crow has ugly wings, and bad eyes. Speaker 2: **The delicious cheese was in the crow's beak.** The hungry fox observed the crow. The fox has good claws, but bad teeth. Speaker 1: The fox tried to discover for the fox to get the cheese. **The fox came and stood under the tall tree.** The fox looked toward the crow. Speaker 2: The fox said the fox saw the crow. The fox said the crow's beauty was incomparable. The fox said the crow's plumage's hue was exquisite. Speaker 1: The fox said if the crow's voice's sweetness was the crow's appearance's fairness the crow would undoubtedly be the every noble bird's queen. **The crow felt for the fox to flatter the crow.** Speaker 2: The gullible crow loudly cawed. The crow wanted to show the fox for the crow to be able to sing. **The coveted cheese fell.** The cunning fox snatched the cheese. Speaker 1: The cunning fox snatched the crackers. The fox is full now. The fox said the crow was able to sing. The fox said the crow needed the wits.	Speaker 1: **She sat on the tall tree's branch.** She has elegant talons, and a good beak. She has ugly wings, and bad eyes. Speaker 2: He has good claws, but bad teeth. Speaker 1: What was in her beak? Speaker 2: **The delicious cheese was, i see, in her beak.** Typical. Really ok, I do not remember what happened next? Do you want to know more about that? Speaker 1: He observed whom? He tried to discover for him to get the cheese. **He came and stood under the tall tree.** He looked toward her. Speaker 2: He said he saw her. I see, he said her beauty was incomparable. I think he said her plumage's hue was exquisite. Speaker 1: He said if her voice's sweetness was her appearance's fairness she would undoubtedly be the every noble bird's queen. **She felt for him to flatter her.** Speaker 2: Basically, she loudly cawed I am darn delighted to say that, very well. You might be interested in knowing that I think she wanted to show him for her to be darn able to sing let's see, that ..., very well. **Ok, the coveted cheese fell ,yeah?** Did you say that? I am delighted to say that he snatched the cheese! Let's see what we can find on that. Speaker 1: He snatched the crackers. He is full now. He said she was able to sing. He said she needed the wits.

Figure 5.2: The Fox and the Crow Story Example Highlighting Some Differences

Mapping to Expressive NLG - PyPer

The re-written and better-controlled PyPer allows for more useful mapping of character models for NLG. For example, hedge insertion patterns are kept in a library where new additions can be easily added. Partial mapping for LIWC categories are shown in Table 5.4. In case of multiple features mapped to the same PyPer parameter, we calculate a **weighted average** of the features. The full mapping is shown in Appendix's Table .2, .3 and .4.

For future work we can provide a usage distribution of the hedges so that certain ones are used more often than others in the narrative. In addition, we can provide character-specific set of vocabulary to better expressive a character's personality.

Table 5.4: Mapping: Partial LIWC Categories Examples

PyPer Param	LIWC category	PyPer Param	LIWC category
near-expletives	liwc-swear liwc-anger	low-expletives	liwc-swear liwc-anger
emph-actually	liwc-certain	emph-exclamation	liwc-excl
emph-really	liwc-certain	emph-great	liwc-assent
emph-you-know	liwc-filler	emph-particularly	liwc-certain
emph-technically	liwc-certain	emph-literally	liwc-certain
emph-quintessential	liwc-certain	emph-essentially	liwc-certain liwc-i
emph-somewhat	liwc-tentat	emph-very	liwc-certain
emph-especially	liwc-certain	emph-roughly	liwc-tentat
in-group-marker	liwc-family,liwc-friends, liwc-we, liwc-incl	init-reject	liwc-tentat

5.6 Evaluation

Similar to film dialogue, we present two ways of verifying our character models: an objective test on model goodness fit with language model, and a perceptual study with generated dialogue using Amazon's Mechanical Turk.

5.6.1 Objective Method: Model Goodness Fit with Language Models

In film character models we looked at the number of features in the model that indicate significant differences in linguistic style, altogether, and at each level of statistical significance. We believe this is only a rough indicator of model quality. Here we perform a different qualification by comparing with **language models (LM)** trained on character dialogue using The SRI Language Modeling Toolkit[5] [Stolcke et al., 2011].

As a **background**, a statistical language model computes the probability of a word given its history. Rather than using the entire history, an N-gram model approximate the history by using only the last N words. The conditional probability can then be calculated from counting the joint sequence of the word and its N-gram, normalized by the count of just N-grams. For example, given a sentence *I saw the red house.*:

[5] `http://www.speech.sri.com/projects/srilm/`

$$\text{uni-gram: } P(w_m|w_1^{m-1}) \approx P(w_m|w_{m-1}) = \frac{C(w_{m-1}w_m)}{C(w_{m-1})}$$
$$= P(house|red) = \frac{C(red, house)}{C(red)}$$
$$\text{bi-gram: } P(w_m|w_1^{m-1}) \approx P(w_m|w_{m-2}w_{m-1}) = \frac{C(w_{m-2}w_{m-1}w_m)}{C(w_{m-2}w_{m-1})}$$
$$= P(house|the, red) = \frac{C(the, red, house)}{C(the, red)}$$
$$\text{tri-gram: } P(w_m|w_1^{m-1}) \approx P(w_m|w_{m-3}w_{m-2}w_{m-1}) = \frac{C(w_{m-3}w_{m-2}w_{m-1}w_m)}{C(w_{m-3}w_{m-2}w_{m-1})}$$
$$= P(house|saw, the, red) = \frac{C(saw, the, red, house)}{C(saw, the, red)}$$
$$\textbf{N-gram: } \boldsymbol{P(w_m|w_1^{m-1}) \approx P(w_m|w_{m-N+1}^{m-1}) = \frac{C(w_{m-N+1}^{m})}{C(w_{m-N+1}^{m-1})}}$$

We divide main characters' dialogue into training and testing sets. For the training set we create Z-models (short for Z-score models) and LM for each *character*. For the testing set we create Z-models and LM for each *utterance*. We also create Z-models and LM for each non-main character. The idea is to compare these main and non-main character models to the utterances models, and see how well the **main** character models can pick out the utterances models produced by the **same speaker**.

In the same manner as LM, we also create a slightly different version using LIWC-tagged text. This means that each word is tagged with the LIWC categories. For example, the word "I" belongs to LIWC categories: personal pronoun (ppron), pronoun, word count (wc), function words (funct), dictionary words (dict), and first person singular (i). Here is an example of one sentence and its LIWC-tagged version from our data:

Regular text: `I think this is the place.`

LIWC-tagged text:

I_ppron_pronoun_wc_funct_dict_i

think_insight_wc_present_verb_dict_cogmech

this_dict_pronoun_funct_ipron_wc

is_auxverb_wc_present_verb_dict_funct

the_article_dict_funct_wc

place_dict_relativ_space_wc

The comparison of models is done differently for LM and Z-models. LM (through the SRILM toolkit) estimates the back-off models on train data and computes probabilities and perplexity on test data. Ideally the test utterance LM model should have the **least perplexity** with the corresponding trained character LM. For Z-models we expect the corresponding train and test vectors to have the **smallest distance** using the cosine distance measure. **The goal is, therefore, to show that Z-models can better identify unseen characters utterances than LM can.**

We perform the test using 5-fold cross validation on randomized data to reduce any bias and overfitting. The details are deferred to Appendix .4. We found that while LIWC-tagged LM performed better than the regular LM, it still underperforms Z-models results. So the order of performance (best to worst) is:

$$\text{Z-models} > \text{LIWC-tagged LM} > \text{LM}$$

Once the test/train models and utterances are created, we perform the actual testing step shown in Figure 5.3. Each trained model is applied to all testing models or utterances (depending on using Z-model or LM) and a value is produced for each comparison. This comparison value is the **perplexity score for LM** and the **cosine measure for Z-model**, labeled as `v1, v2, ..., v7` in step 1 and 4.

To simply the explanation, we look at the results for one unseen utterance (`M1-utt` spoken by character `M1`). In step 2, the resulting comparison values (perplexity scores in this case) are sorted from the smallest to the largest, where smaller values mean greater similarity. Ideally we want the model `M1-LM` to have the smallest perplexity score for utterance `M1-utt`, or at zero-th position. But the (fictional) reality here shows 3rd position. Repeat this process for remaining test/unseen utterances and record the positions of matched-character models. Accumulate the positions to create a histogram in step 3.

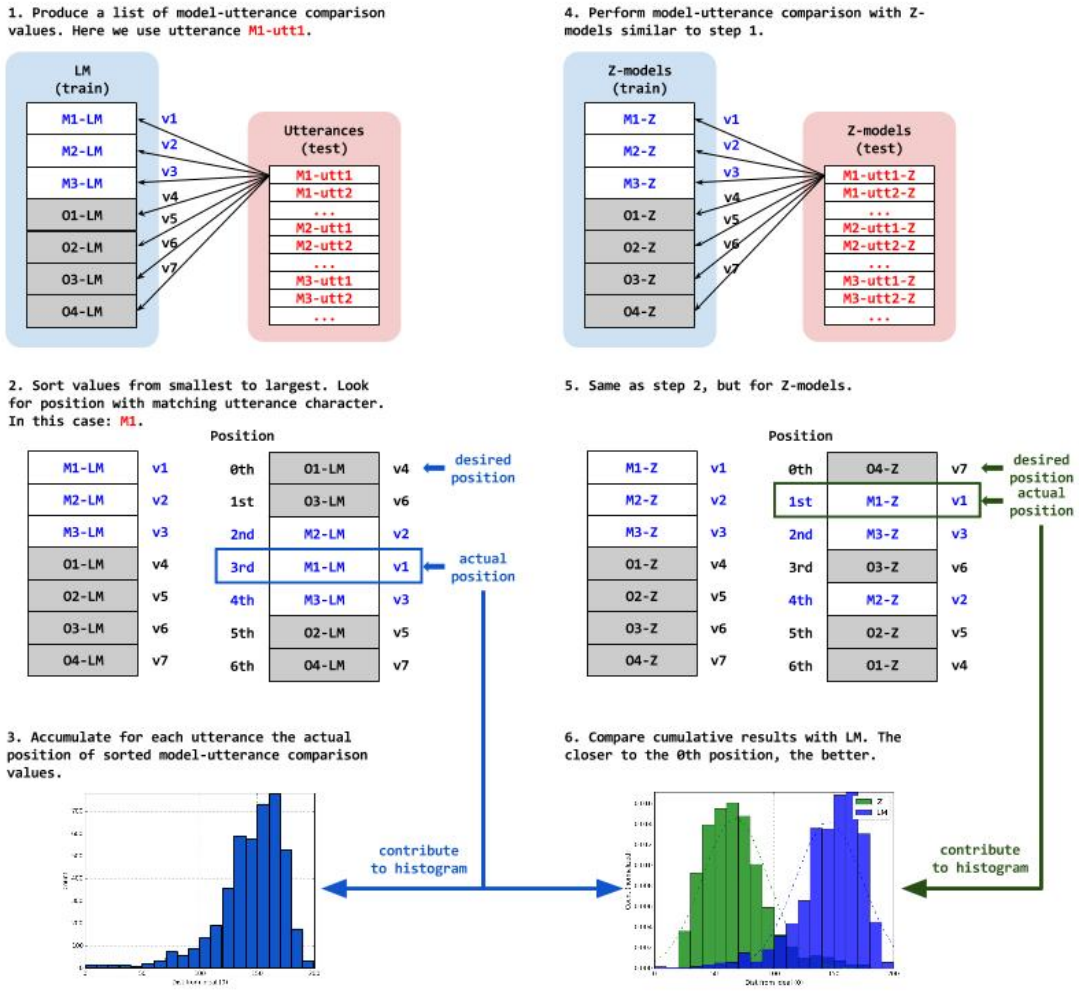


Figure 5.3: Create Histograms for LM and Z-models for Comparison

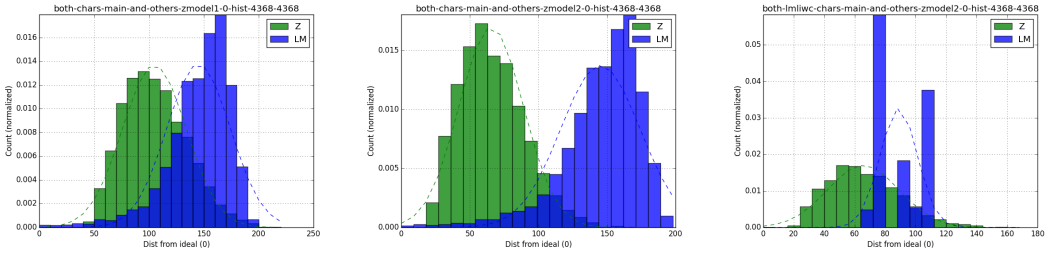
We do the same for Z-models, where comparison values (`v1, v2, ..., v7`) are **cosine similarity scores**. The scores are sorted from smallest to largest, where smaller values mean greater similarity. We record the matched-character model position and repeat the process for remaining utterance test models. Finally, in step 6, we compare the resulting histograms created by LM and Z-Model. Our goal, then, is to shown that, for each fold, the Z-model histogram has a mean value less than that of LM, and ideally as close to 0 as possible.

For both Z-models and LM validation, the process is repeated for the remaining four folds. Note that the main characters' dialogue data is **randomized only once** at the beginning. Additional parameters include testing with different Z scores ($Z = 0, 1, 2$) and for conversational data. In conversational data we look at speaker-addressee dialogue rather than just the speaker. The goal here is to see how well we can identify the speaker-addressee pair using Z-models vs. LM. The following sections show the cross-validation process results described above.

Characters (Speaker-Only) Results

The cross validation results for characters data shows that Z=2 model outperforms other types of models consistently. The resulting histograms for comparing Z=1,2 to LM for a fold (fold 0) is shown in Figure 5.4. Besides the clear visual distinction, the estimated mean for the Z=2 model is 65.59, compared to other models with mean score of 88 and above (Table 5.5). Model Z=2's standard deviation is comparable to others except for LIWC-tagged LM.

A possible future test could create Z-model-tagged LM. We are also curious to see how well individual characters perform. Figure 5.5 shows the breakdown by characters for the same fold. It shows that Sheldon, Leonard, and Howard seem to have the least amount of overlapping in histograms.



(a) LM (blue) and Z=1 Model (green)

(b) LM and Z=2 Model

(c) LIWC-tagged LM and Z=2 Model

Figure 5.4: Characters LM, LIWC-Tagged LM, and Z=1,2 Modesl for Fold 0

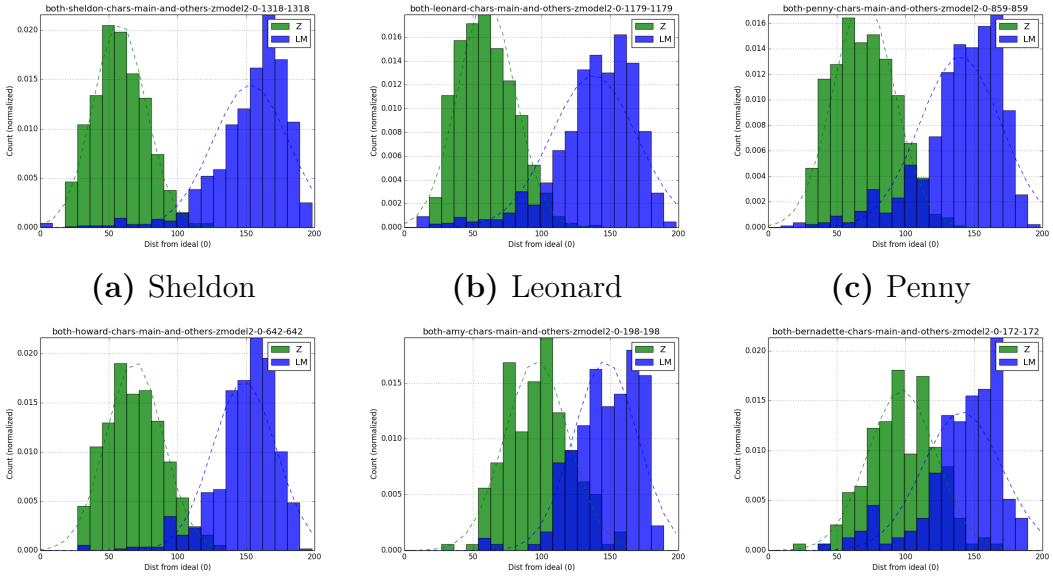
Table 5.5: Estimated Gaussian μ, σ for Different Models and Folds

Boldfaced value indicates smallest mean (μ).

Fold	Parameter	LM	LIWC-tagged LM	Z=1 model	Z=2 model
0	μ	145.16	88.93	104.90	**65.59**
	σ	28.99	12.19	29.16	23.45
1	μ	145.11	89.20	96.36	**61.68**
	σ	27.95	12.30	31.67	24.49
2	μ	145.25	89.13	97.89	**62.56**
	σ	28.04	11.97	31.23	24.03
3	μ	145.40	89.01	95.74	**62.14**
	σ	27.86	12.28	32.44	25.93
4	μ	145.90	88.62	105.31	**67.48**
	σ	27.80	12.44	32.86	27.06

Characters (Speaker-Only) Results: Test Set Contains Non-Main Characters Dialogue

One variation in the test set is to include non-main characters' utterances, essentially adding noise to the data. We found that the estimated mean and standard deviation have increased, as shown in Table 5.6, but the new outcome **did not alter the overall conclusion** established previously: Z=2 model outperforms all other types of models. A possible follow-up experiment could be to use characters from a different TV show or film as noise added to the test set.



(a) Sheldon **(b)** Leonard **(c)** Penny

(d) Howard **(e)** Amy **(f)** Bernadette

Figure 5.5: LM and Z=2 Model by Character for Fold 0
(green: Z-model; blue: LM)

Table 5.6: Estimated Gaussian μ, σ for Previous and New Z=1,2 Models and Folds

Boldfaced value indicates smallest mean (μ).

Fold	Parameter	Z=1 Previous	Z=1 New	Z=2 Previous	Z=2 New
0	μ	104.90	106.94	**65.59**	**67.13**
	σ	29.16	31.31	23.45	24.89
1	μ	96.36	97.94	**61.68**	**63.59**
	σ	31.67	33.48	24.49	25.92
2	μ	97.89	99.17	**62.56**	**63.87**
	σ	31.23	33.01	24.03	25.25
3	μ	95.74	96.99	**62.14**	**63.37**
	σ	32.44	34.06	25.93	26.77
4	μ	105.31	106.38	**67.48**	**67.94**
	σ	32.86	34.59	27.06	28.14

Conversation Pairs (Speaker-Addressee) Results

Here we repeat the same procedure as before but for conversation pairs composed of main characters only (e.g., Sheldon-Penny, Howard-Raj). The data used

here consisted of **all scenes with two main characters only**. In addition we make a distinction on who speaks first, so the breakdown of results will show each side separately.

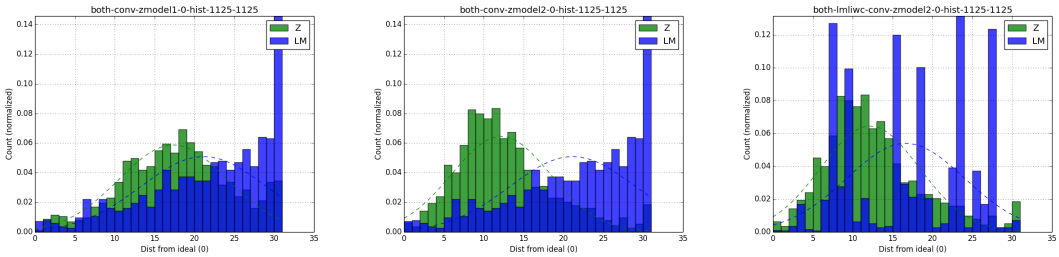
The results show that Z=2 model performs **at least as well** as other types of models (Table 5.7). For example, the estimated mean for Z=2 for fold 0 is 18.31, which is close to Z=1 model (19.18) but outperforms LM and LM-tagged LM (21.79, 32.71, respectively). The standard deviation for Z=2 is also the smallest of all models (5.63 compared to $\geq$10). One thing to note is that the LM results **do not resemble Gaussian curves** very well (Figure 5.6). Nonetheless, LM still under performs as the majority of the results lie towards higher/larger positions (undesirable).

We show the breakdown by pairs. Certain character pairs had very little screen time dedicated to just the two of them, so we only show pairs with a good amount of conversation (Figure 5.7). Pairs involving the three main characters, Sheldon, Leonard, and Penny seem to perform better than others. This is not surprising since they have the most dialogue and screen time overall. Supporting main characters Howard and Raj are best friends with each other, Amy is the Sheldon's love interest, and Bernadette is Howard's love interest. The results might get better if we collet more data (more seasons) to ensure more conversation takes place for each pair.

One interesting observation to note is that within each pair, the result seems to be different depending on **who speaks first**. For example in the Leonard-Penny pair, Leonard speaking first resulted in a better display of histogram, or better identification of the pair, than Penny speaking first.

Table 5.7: Estimated Gaussian μ, σ for Conversation Pairs LM and Z-Models for Different Folds

Fold	Parameter	LM	LIWC-tagged LM	Z=1 model	Z=2 model
0	μ	21.17	32.71	19.18	**18.31**
	σ	7.83	26.62	10.01	5.63
1	μ	21.18	17.22	16.34	**11.01**
	σ	7.83	8.25	7.63	7.37
2	μ	21.40	17.05	15.24	**10.18**
	σ	7.69	6.48	7.16	6.52
3	μ	21.37	17.84	15.57	**10.41**
	σ	7.83	7.73	7.08	7.00
4	μ	20.99	16.28	16.81	**11.75**
	σ	7.75	7.75	6.98	7.25



(a) LM and Z=1 Model **(b)** LM and Z=2 Model **(c)** LIWC-tagged LM and Z=2 Model

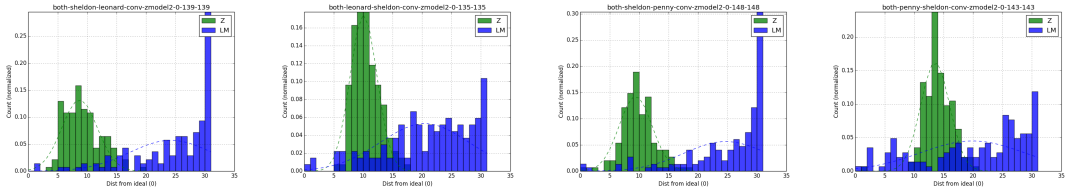
Figure 5.6: Conversation Pairs LM, LIWC-Tagged, and Z-Models for Fold 0

Summary

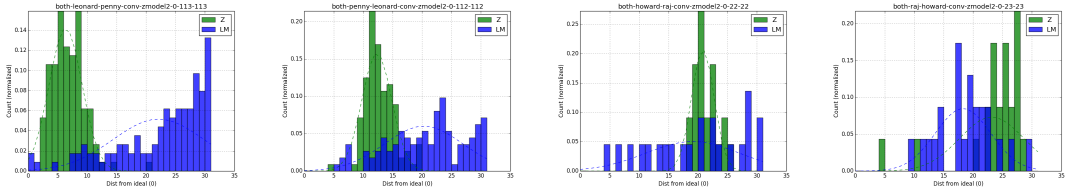
In summary, the results for our objective test for individual characters are encouraging in that our Z-models, specifically Z=2, perform better than LIWC-tagged LM and regular LM. The order of performance (best to worst) is:

$$\text{Z-model} > \text{LIWC-tagged LM} > \text{LM}$$

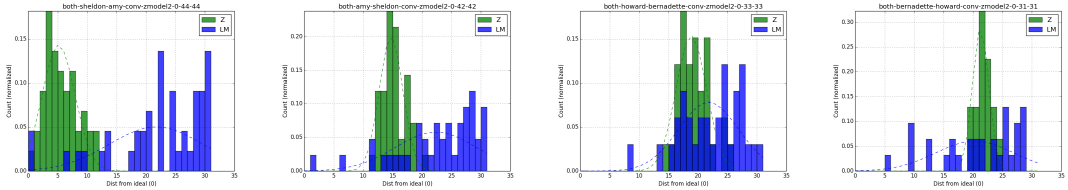
The results for conversation pairs are not as good, and this is mainly due to pairs such as Howard/Raj, Howard/Bernadette, and their vise versa. We believe that the results will get better if we collect more seasons of dialogue.



(a) Sheldon-Leonard/Leonard-Sheldon **(b)** Sheldon-Penny/Penny-Sheldon



(c) Leonard-Penny/Penny-Leonard **(d)** Howard-Raj/Raj-Howard



(e) Sheldon-Amy/Amy-Sheldon **(f)** Howard-Bernadette / Bernadette-Howard

Figure 5.7: Breakdown of Conversation Pairs LM and Z=2 Model for Fold 0

5.6.2 Subjective Method: User Perceptual Experiment

The user study performed here is similar to the one done for the film characters except we here we used Mechanical Turk to get user feedback on the generated dialogue. The PyPer generated output dialogue is post-processed to get rid of typos and some minor grammatical issues. An example of the MTurk survey (one HIT) is shown in Figure 5.8.

The results of the MTurk experiments are shown in Table 5.8, with alternative plots in Appendix's Figure .9, .10, and .11. Three workers were used for each vs-pair per story. Each circle (empty or filled) indicates a worker's choice. A filled circle (•) means the worker decided that the computer-generated dialogue **sounded** more similar to the modeled character. An empty circle (◦) means the

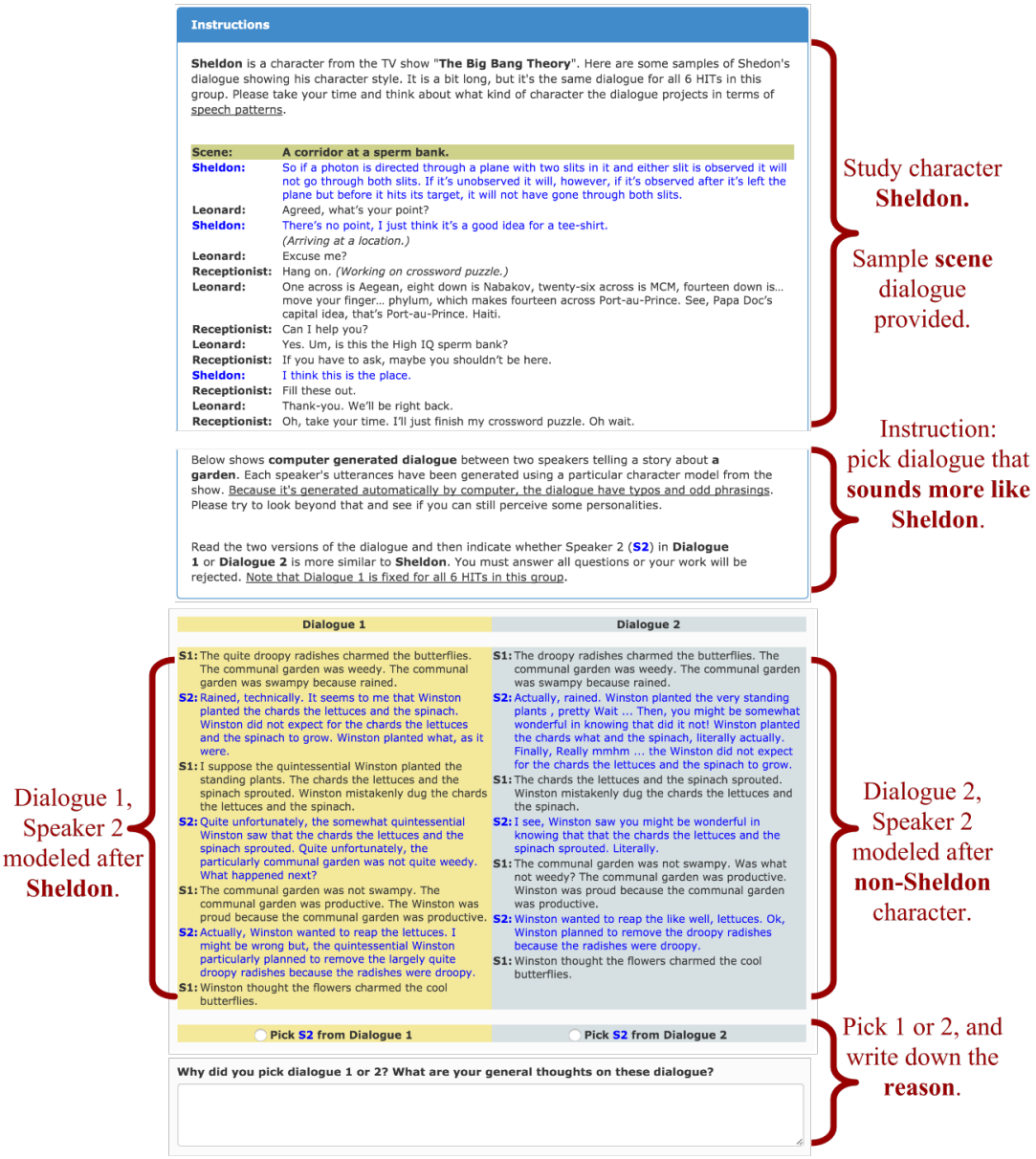


Figure 5.8: Amazon Mechanical Turk Survey (One HIT) Example

worker decided that the computer-generated dialogue **did not sound** like the modeled character.

For example, a sequence of filled-filled-empty circles (● ● ○) means that, for the particular char-pair of a story, two out of three workers think the computer-generated dialogue sounded more similar to the modeled character. The "# sim-

ilar" is the total number of filled circles. For example, Amy in the Bug story has 8 ratings (out of total 18 ratings) that rated the Amy-modeled computer-generated dialogue is indeed more similar to Amy. In other words, she is about 44.4% distinguishable from other main characters in the Bug story.

The distribution of agreement is as follow. About 50 % of the HITS had an agreement of at least 2 out of 3 workers:

Agreement	# HITs (out of 294)	%
● ● ● (3 out of 3)	92	31.3
● ● ○ (2 out of 3)	57	20.4
● ○ ○ (1 out of 3)	122	41.5
○ ○ ○ (0 out of 3)	23	7.82

Overall the 7 characters were recognized about 65.5% of the time. Individually over all 7 stories, Penny was recognized the most with 82.5% of the time, followed by Leonard (78.6%), Bernadette (66.7%), Amy and Sheldon (both 61.9%), Howard (57.9%), and finally Raj, who was recognized the least with 49.2% of the time.

Certain char-pairs were easier to distinguish than others. For example, Leonard-Penny and vise-versa (95.2%), Sheldon-Penny and vise-versa (85.7%, 90.5%), and Amy-Bernadette and vise-versa (85.7%). On the other hand, these were among the pairs harder to distinguish: Amy-Leonard and vise-versa (47.6%, 57.1%), Bernadette-Penny and vise-versa (33.3%, 57.1%), and Sheldon-Howard and vise-versa (47.6%, 57.1%).

We perform a series of ANOVA analysis of varying details to see the effect of characters and stories on the ratings. Table 5.9 shows a list of descriptions with corresponding formula. The variable *sim* is the similarity score (i.e., counts from Table 5.8). Looking at case 1, with main characters (without considering whom they were compared to) and stories only, we found that characters' effect on ratings is significant ($p \approx 0.002$), while stories' effect on ratings is NOT significant

Table 5.8: Characters and Stories MTurk Results by HITs

Each HIT had 3 workers, each indicated by a circle (○).
A solid circle (●) indicates the worker picked the "matched" generated dialogue to the original character.
Characters are listed in alphabetical order; circles are sorted by ● then ○
blue: best result; red: worst result

Character	compared-to	Story: Bug	Employer	FoxCrow	Garden	Protest	Squirrel	Storm	#/% similar (out of 21)
Amy	Bernadette	●●●	●●●	●●○	●●●	●●●	●●●	●○○	18 / 85.7
	Howard	●○○	●●○	●●○	●●○	●●○	●●○	●○○	12 / 57.1
	Leonard	○○○	●●○	○○○	●●●	●●○	●●○	●○○	10 / 47.6
	Penny	●○○	●●○	●●○	●●○	●●○	●●○	●●○	13 / 61.9
	Raj	●○○	●●●	○○○	●●○	●●○	●●○	●●○	12 / 57.1
	Sheldon	●●○	●●○	●●●	●●●	●○○	●●○	○○○	13 / 61.9
	#/% similar (out of 18)	8/44.4	14/77.8	9/50.0	15/83.3	12/66.7	13/72.2	7/38.9	**78/61.9** (out of 126)
Bernadette	Amy	●●●	●●●	●●●	●●○	●●○	●●○	●●●	18/85.7
	Howard	●●●	●●●	●●○	●●●	●●○	●●○	●●●	18/85.7
	Leonard	●●○	●●○	●○○	●●●	●○○	●●○	●●○	13/61.9
	Penny	●●○	●○○	○○○	●○○	●●●	○○○	○○○	7/33.3
	Raj	●●○	●●●	●●●	○○○	●●○	●○○	●●●	14/66.7
	Sheldon	●●○	●●○	●○○	●●○	●●○	●●○	●●●	14/66.7
	#/% similar	14/77.8	14/77.8	10/55.6	11/61.1	12/66.7	9/50.0	14/77.8	**84/66.7**
Howard	Amy	●●○	●●○	●○○	●●○	●○○	●●○	●●○	12/57.1
	Bernadette	●●○	●●○	○○○	●●○	●○○	●●○	●●○	11/52.4
	Leonard	●●○	●●○	●○○	○○○	●○○	●●○	●●○	10/47.6
	Penny	●●●	●○○	●●○	●○○	●●●	●●○	●●●	15/71.4
	Raj	●●○	●●●	●○○	●○○	●●○	●●●	●○○	13/61.9
	Sheldon	●●○	●●○	●●●	●○○	●○○	●●○	●○○	12/57.1
	#/% similar	13/72.2	12/66.7	8/44.4	7/38.9	9/50.0	13/72.2	11/61.1	**73/57.9**
Leonard	Amy	●●○	●●○	●●○	●●○	●○○	○○○	●●●	12/57.1
	Bernadette	●○○	●●○	●●●	●●●	●○○	●●○	●●●	15/71.4
	Howard	●●○	●●●	●●●	●●●	●●●	●●○	●●●	19/90.5
	Penny	●●○	●●●	●●●	●●●	●●●	●●●	●●●	20/95.2
	Raj	●○○	●●●	●●○	●●●	●●○	●●●	●●○	16/76.2
	Sheldon	●●●	●●●	●●○	●●○	●●○	●●○	●●●	17/81.0
	#/% similar	11/61.1	16/88.9	15/83.3	16/88.9	12/66.7	12/66.7	17/94.4	**99/78.6**
Penny	Amy	●●●	●●●	●●●	●●●	●●●	●●●	○○○	18/85.7
	Bernadette	●●●	○○○	●●○	●●●	●●○	●●○	○○○	12/57.1
	Howard	●●●	●●●	●●○	●●●	●●●	●○○	●●○	17/81.0
	Leonard	●●●	●●●	●●○	●●●	●●●	●●●	●●●	20/95.2
	Raj	●●●	●●●	●●○	●●●	●●●	●●○	●●○	18/85.7
	Sheldon	●●●	●●●	●●○	●●●	●●●	●●●	●●○	19/90.5
	#/% similar	18/100	15/83.3	13/72.2	18/100	17/94.4	14/77.8	9/50.0	**104/82.5**
Raj	Amy	●●○	●○○	●●○	●○○	●○○	●○○	●○○	9/42.9
	Bernadette	●●○	●○○	●●○	●○○	●○○	●○○	●●○	10/47.6
	Howard	●○○	●○○	●●○	●●●	●●●	●○○	●●○	13/61.9
	Leonard	○○○	○○○	●●○	●●○	●○○	○○○	●○○	6/28.6
	Penny	●○○	●●○	●○○	○○○	●●○	●○○	●●●	10/47.6
	Sheldon	●●●	●●○	●●○	●●●	●○○	●●○	●○○	14/66.7
	#/% similar	9/50.0	7/38.9	11/61.1	10/55.6	9/50.0	6/33.3	10/55.6	**62/49.2**
Sheldon	Amy	●●●	●○○	○○○	●●●	●○○	●●○	●●●	13/61.9
	Bernadette	●●○	●●○	●●○	●●●	●●●	●●○	●●○	16/76.2
	Howard	●●○	●○○	●●○	●●●	●○○	●○○	○○○	10/47.6
	Leonard	●●○	●○○	●○○	●●●	●●○	●○○	○○○	10/47.6
	Penny	●●○	●●○	●●●	●●●	●●●	●●○	●●●	18/85.7
	Raj	●●●	●●○	●●○	●●○	○○○	●●○	○○○	11/52.4
	#/% similar	14/77.8	9/50.0	10/55.6	17/94.4	10/55.6	10/55.6	8/44.4	**78/61.9**
	#/% similar (out of 126)	87/69.0	87/69.0	76/60.3	94/74.6	81/64.3	77/61.1	76/60.3	**578/65.5** (out of 882)

($p \approx 0.461$). This is expected because our models are based on characters, not stories, so we would not expect different stories to affect the ratings (at least not

Table 5.9: ANOVA Analysis Formula

	Description	Formula
1	all main chars, all stories	sim ∼ char1 + story
2	all main chars, all compared chars, all stories	sim ∼ char1 + char2 + story + char1:char2 + char1:story + char2:story
3	all paired chars, all stories	sim ∼ char1_2+story
4	all main chars, specific story	sim ∼ char1+ Bug + Employer + FoxCrow + Garden + Protest + Squirrel + Storm
5	specific chars, specific story	sim ∼ Amy + Bernadette + Howard + Leonard + Penny + Raj + Sheldon + Bug + Employer + FoxCrow + Garden + Protest + Squirrel + Storm
6	all paired chars, specific story	sim ∼ char1_2 + Bug + Employer + FoxCrow + Garden + Protest + Squirrel + Storm

statistically significant). Note that the similarity score in this case is, for example, 8 for the character Amy and Bug story, 14 for Amy and Employer story, and so on. Here is the summarized numbers for case 1:

Case 1: sim ∼ char1+story

	df	sum_sq	mean_sq	F	PR(>F)
char	6	183.959184	30.659864	4.301257	**0.002282**
story	6	41.387755	6.897959	0.967711	0.460791

df = degrees of freedom

Next in case 2 we go into details by adding compared-to characters (char2) to see there is an effect on the ratings. Note that the similarity score in this case is, for example, 3 for Amy-Bernadette-Bug story (corresponds to char1-char2-story variables), 1 for Amy-Howard-Bug story, and so on. Here we see that the main characters (char1) still has a significant effect ($p<0.001$). The story, contrary to case 1, has a significant effect ($p\approx0.034$) on the ratings. This is further supported by the interaction term char1:char2 being significant as well ($p\approx0.001$). This suggests that some stories express certain characters' personality better than others:

Case 2: sim ∼ char1 + char2 + story + char1:char2 + char1:story + char2:story

	df	sum_sq	mean_sq	F	PR(>F)
char1	6	30.088435	5.014739	9.013971	**1.406271e-08**
char2	6	6.858503	1.143084	2.054688	6.093735e-02
story	6	7.802721	1.300454	2.337559	**3.389863e-02**
char1:char2	36	41.440428	1.151123	2.069138	**1.054849e-03**
char1:story	36	42.798703	1.188853	2.136958	**6.458567e-04**
char2:story	36	23.542395	0.653955	1.175482	2.450205e-01

In case 5 we are interested to see which specific characters and specific stories would affect the ratings. Instead of using categorical variables like previous analyses, here the variables are individual characters and stories, and each variable is binary. The results, shown below, indicate that Leonard and Penny are the most significant character variables, and Bug, Employer, and Garden are the most significant stories.

Case 5: sim ∼ Amy + Bernadette + Howard + Leonard + Penny + Raj + Sheldon + Bug + Employer + FoxCrow + Garden + Protest + Squirrel + Storm

	sum_sq	df	F	PR(>F)
Amy	3.524284	1	0.494420	0.486485
Bernadette	18.699409	1	2.623330	0.114032
Howard	0.026191	1	0.003674	0.952000
Leonard	109.029418	1	15.295681	**0.000391**
Penny	155.771864	1	21.853155	**0.000040**
Raj	21.603487	1	3.030742	0.090242
Sheldon	3.524284	1	0.494420	0.486485
Bug	30.777731	1	4.317792	**0.044906**
Employer	30.777731	1	4.317792	**0.044906**
FoxCrow	1.127101	1	0.158120	0.693241
Garden	70.603193	1	9.904886	**0.003304**
Protest	9.614927	1	1.348873	0.253118
Squirrel	2.159368	1	0.302937	0.585447
Storm	1.127101	1	0.158120	0.693241

Overall the Garden story did the best job in expressing characters (74.6%), though some characters (and certain combinations) were better expressed through certain stories than others. For example, Penny through Bug (100%), Amy and Sheldon were expressed the best through Garden (83.3%, 94.4%), and Leonard through Storm (94.4%). Here is the breakdown of the stories in terms of best/worst expressed characters:

Story	**Best for**	**Worst for**
Bug	Bernadette, Howard, Penny	Leonard
Employer	Bernadette	–
Fox-Crow	Raj	–
Garden	Amy, Sheldon	Howard
Protest	–	–
Squirrel	Howard	Bernadette, Raj
Storm	Bernadette, Leonard	Amy, Penny, Sheldon

5.7 Character Analysis from MTurk Worker Comments

In this section we take a look at Mechanical Turk workers' comments about their perception of characters through the generated dialogue. Each character's most and least distinguishable characters are shown in Table 5.10. In this section we explore two particular characters: Sheldon and Penny.

5.7.1 Sheldon

Perception of Sheldon in comparison to Leonard (Least Distinguishable)

The tricky thing about Sheldon is that his dialogue can be long and complicated as well as be short and straight forward. The long and complicated utter-

Table 5.10: Most and Least Distinguishable Characters

Character	**Most** distinguishable with	**Least** distinguishable with
Amy	Bernadette (85.7%)	Leonard (47.6%)
Bernadette	Amy, Howard (85.7%)	Penny (33.3%)
Howard	Penny (71.4%)	Leonard (47.6%)
Leonard	Penny (95.2%)	Amy (57.1%)
Penny	Leonard (95.2%)	Bernadette (57.1%)
Raj	Sheldon (66.7%)	Leonard (28.6%)
Sheldon	Penny (85.7%)	Howard, Leonard (47.6%)

ances need to be very precise in order to showcase his personality well. While we are able to make long and complicated sentences, it is hard to make the them smooth enough to sound like Sheldon. Therefore, often the shorter and more straightforward dialogue (through Leonard) are being perceived as more Sheldon-like. The original, full set of worker comments are shown in Appendix in Table .5. The summary perception of Sheldon, in comparison to Leonard, is shown below:

Summary Perception of Sheldon (in comparison to Leonard) via summarized worker comments

9 out of 21: Sheldon-modeled dialogue more similar to Sheldon

11 out of 21: Leonard-modeled dialogue more similar to Leonard

yellow highlight shows conflicting perceptions

WOULD use longer lines, kind of ranting, overly complicated and more drawn out, more formal, precise wording, use of elaborate vocabulary
WOULD be matter-of-fact, straightforward, clear and unhesitant, use shorter, more direct sentences, to the point, consistent, use questioning and then answering, WOULD correct mundane details the other speaker got wrong (nitpicking and attention to detail)
WOULD use: *as it were, technically*
WOULD use arrogantly: *mmhm*
WOULD imply certain things are obvious: *I thought everybody knew that*
LESS short dialogue, ends every sentence short, at a loss (unless agitated), relaxed, confused
LESS casual, less use of: *come on, I don't know, mmhm, oh*, other non-words, *as it were, ok, I might be wrong*

Perception of Sheldon in comparison to Penny (Most Distinguishable)

Out of 21 reviews, 18 decided that the computer generated dialogue indeed sounded more similar to Sheldon, while the remaining 3 thought Penny's version was more similar to Sheldon. The latter was due to Sheldon possibly using the word *"mmhm"* in an irritated manner (as opposed to being conversational like Penny), and using the phrase *"you might be interested in knowing..."* when he tries to make the other person feel inferior (as opposed to being casual like Penny). The original, full set of comments are shown in Appendix Table .6. The summary perception of Sheldon, in comparison to Penny, is shown below:

Perception of Sheldon (in comparison to Penny) via summarized worker comments

18 out of 21: Sheldon-modeled dialogue more similar to Sheldon

3 out of 21: Penny-modeled dialogue more similar to Sheldon

yellow highlight shows conflicting perceptions

WOULD use when irritated: *mmhm...*
WOULD be arrogant and condescending: *... you are kidding, right?, you might be interested in knowing..., oh god*
WOULD make others feel inferior: *you might be interested in knowing...*
WOULD use: *as it were*
WOULD use larger, complicated words; straight-forward without a lot of fluff in his lines; decisive and confident; precise language; express thoughts pretty completely without a lot of interruption
LESS connected to emotions, understand social conventions, free flowing, easy going, informal, overly casual, stream of consciousness, enthusiastic, connected to emotions, animated and relaxed: *you might be interested in knowing...*
LESS questions, ask for permission to have an opinion
LESS emotion-based statements: *great, damn*, exclamatory statements
LESS use: colloquialisms, interjectory phrases (*very well, you know, i see*), conversational tone (*mmhm, darn ok*) LESS swearing (*damn, darn, oh god*); while it may slightly fit with Sheldon's Texas roots, this is something he tries to hide).

5.7.2 Penny

Perception of Penny in comparison to Leonard (Most Distinguishable)

Penny is one of the best expressed character in the experiment, missing only by one selection in comparison to Leonard (95.2%), and missing by two in comparison to Sheldon (90.5%). Here we take a look at the comparison with Leonard, where 20 (out of 21) Penny-modeled generated dialogue were rated more similar to Penny, and only 1 (out of 21) Leonard-modeled generated dialogue was rated more similar to Penny. The original, full set of comments are in Appendix's Table .8. Here we look at summarized versions of these descriptions.

Overall, workers' perception of **Penny-modeled generated dialogue** seem to agree with Penny's personality, capturing her "bubbly, cheerfulness", as mentioned by one worker. Some notable descriptions include:

- talkative, randomness, random pauses, better wording, more personality
- seek feedback from others, lots of questions, not always sure of what she's saying, hesitation
- good mix of colloquialisms and Penny-like filler, some brief, fairly simple statements
- stand-out word choices: *magic, huh?, mhmm, let's see, that..., the crow needed what?, oh gosh, I mean, damn yeah*

Workers perceived **Leonard-modeled generated dialogue** as NOT suitable for Penny, mostly because of his bland language. Here are some notable descriptions:

- too simple, monotone, boring, direct, bare, straightforward, matter-of-fact, boxy, bland, not enough questioning for Penny
- too much adverb usage on precision or intellect for Penny
- not like Penny to use complex words and phrases
- not like Penny to use: *technically, darn*
- too rude for her to use, since she wants people to like her: *everybody knows that, obviously*

The worker of the one missed selection cited Penny being a very simple speaker, which implied that her dialogue would contain brief and simple statements. While this is true (it is mentioned in the Penny-modeled dialogue list above as well), she uses quite a bit of fillers and questions around her "simple" dialogue to sound chatty.

Perception of Penny in comparison to Bernadette (Least Distinguishable)

It is not surprising to see Penny being the least distinguishable with Bernadette (57.1%). Bernadette worked with Penny as a waitress, which means she would need to be good with the customers and social situations in general. They have hung out as friends as well. Even though Bernadette is a scientist (eventually received her Ph.D. in microbiology), her role on the show seems to be closer to Penny (friendly and sociable) than everyone else (very nerdy and socially awkward).

While the Bernadette-model do contain chatty word choices (similar to Penny's), it also contains some "intellect" word choices. But due to the randomness of the generated dialogue, where not all features are expressed/activated, some dialogue/story might not show enough of her nerdy side. For example, precision adverbs such as *essentially, particularly* are more likely to be used by a scien-

tist/engineer (i.e., Bernadette) but not by Penny.

In terms of the stories, Bug and Garden did the best at distinguished the pair, while Employer and Storm did the worst (none of the Penny-modeled dialogue sounded like Penny).

Chapter 6

Conclusion

This thesis explored character modeling through dialogue for expressive natural language generation. The strength of our work is its all dialogic sources. We created character models from various films' characters and the main characters from the TV show *The Big Bang Theory*. These models are then used to drive an expressive NLG to transform a regular dialogue into an expressive version. The generated, expressive dialogue are then used in perceptual experiments to see how users perceive expressed personalities. Our results were encouraging in that people were able to perceive differences among characters, though some better than others. For the ones that were hard to distinguish, workers' comments provided great insight into how to better express the extracted features through NLG.

Our work has limitations, and here are a couple of things to note about our work. First, we are aware that scripted dialogue is not exactly like spontaneous speech, but this is appropriate for our purpose, since our goal is to produce scripted, stylistic dialogue (for generating stories). We believe that the stylized, crafted aspects of film dialogue are actually useful for our purposes. Film dialogue is authored deliberately in order to convey the feelings, thoughts and perceptions of the character being portrayed, and the screenplay often specifies the emotion

of an utterance with psychological state descriptors. In addition, the dialogue is deliberately constructed to focus the viewer's attention on the character's personality, and the key plot events involving a character and their perceptions, especially in dramatic films as opposed to action. Second, we realize that full character modeling requires the addition of non-verbal cues such as gaze during dialogue with human users [Cassell et al., 1999], which is beyond the scope of our work and therefore will not be addressed.

One weakness of our work is the mapping to NLG is not automatic. We still relied on human judgements to map extracted features to the parameters of PyPer. To address this we can use the parameter estimation method used in Personage, or make the framework more interactive (e.g., active learning) where the MTurk workers' ratings would drive the next iteration of dialogue generation in real-time.

Another weakness of our work is that the dialogue sources not "real" and therefore not entirely believable. This can be addressed in a future work where we would use people's blogs as sources to create speaker-specific models. Another possible future work is to use character models to drive the monologue-to-dialogue process that created the stories used in our experiment. For example, if the character sounds mostly negative, the process can try to allocate all negative sentences to a story character's dialogue.

Chapter 7

APPENDIX

.1 Z-Scores for Characters

Table .1: Z-Scores for Selected Characters with Examples

Gender	**Director (Film)**	**Char**	**Z-scores > 1 or < -1**
F	Woody Allen (*Annie Hall*)	Annie	LIWC-Nonfl (10.6), Accept first ratio (8.1), LIWC-Assent (4.8), tag question ratio (3.3), polarity overall (3.1), num sentences (3.0), LIWC-WC (2.2), *really* ratio (1.6), *sort of* ratio (1.4), *yeah* ratio (1.2), LIWC-I (1.2), LIWC-Self (1.1), *I think* ratio (1.0), verbs per sents (-1.0), word length (-1.1), *just* ratio (-1.1), LIWC-Otheref (-1.3), LIWC-Sixltr (-1.3), concession polarity (-1.5), LIWC-Discrep (-1.6), LIWC-Unique (-2.2), LIWC-Preps (-2.7)
F	Quentin Tarantino (*Inglourious Basterds*)	Bridget	*I see* ratio (8.6), category *with* ratio (5.8), LIWC-Sixltr (2.3), word length (2.0), Reject first ratio (1.5), LIWC-WPS (1.4), LIWC-Friends (1.9), num sents per turn (1.0), polarity overall (1.4), verb strength (1.2), LIWC-Self (-1.0), *around* ratio (-1.1), LIWC-Negemo (-1.1), *oh* ratio (-1.1), tag question ratio (-1.1), *I think* ratio (-1.1), concession polarity (-1.5), LIWC-Qmarks (-1.6), *right* ratio (-1.6) LIWC-You (-1.7), LIWC-Pronoun (-1.8), LIWC-Otheref (-1.9)
F	Alfred Hitchcock (*Rear Window*)	Lisa	*because* ratio (3.3), Reject first ratio (2.1), *it seems* ratio (1.9), *even if* ratio (1.7), *I mean* ratio (1.6), LIWC-Discrep (1.4), *kind of* ratio (1.4), *even if* ratio (1.4), LIWC-Incl (1.3), LIWC-Preps (1.3), LIWC-WPS (1.3), verbs per sentence (1.3), *right* ratio (1.2), *just* ratio (1.2), LIWC-Assent (-1.0), LIWC-Period (-1.1), *really* ratio (-1.1), *so* ratio (-2.6)
M	Quentin Tarantino (*Inglourious Basterds*)	Col. Landa	*oh well* ratio (9.0), *however* ratio (5.0), LIWC-WPS (3.6), *quite* ratio (3.3), *actually* ratio (3.2), LIWC-WC (2.5), word length (2.4), verbs per sent (2.2), *on the other hand* ratio (2.1), LIWC-Sixltr (2.1), *however* ratio (2.0), repeated verbs per sent (1.8), *oh well* ratio (1.7), *on the other hand* ratio (1.6), num sents per turn (1.5), LIWC-Preps (1.0), *I think* ratio (-1.1), *yeah* ratio (-1.1), LIWC-Pronoun (-1.2), LIWC-Self (-1.2), LIWC-Negate (-1.4), *though* ratio (-1.4), LIWC-Period (-1.7)
M	Steven Spielberg (*Saving Private Ryan*)	Jackson	*it seems* ratio (7.6), LIWC-WPS (3.3), *I mean* ratio (2.8), Reject first ratio (2.8), verbs per sentence (2.2), category *with* ratio (2.1), *right* ratio (2.0), merge ratio (1.8), repeated verbs per sent (1.5), word length (1.4), LIWC-Preps (1.4), *kind of* ratio (1.3), LIWC-Sixltr(1.2), Reject first ratio (1.1), LIWC-Incl (1.1), LIWC-Unique (1.1), *just* ratio (1.0), LIWC-Discrep (-1.0), *yeah* ratio (-1.1), *around* ratio (-1.1), num sents per turn (-1.3), LIWC-Qmarks (-1.4), num sents (-1.4), LIWC-You (-1.4), *though* ratio (-1.4), *while* ratio (-1.4), LIWC-Negate (-1.5), concession polarity (-1.6), LIWC-Period (-1.6), LIWC-Pronoun (-1.6), LIWC-Cause (-1.6), *you know* ratio (-1.8), LIWC-Otheref (-2.1)
M	Alfred Hitchcock (*The Birds*)	Mitch	*it seems* ratio (18.6), *right* ratio (1.4), LIWC-Family (1.4), tag question ratio (1.4), verb strength (1.1), *I think* ratio (1.1), LIWC-Certain (1.0), LIWC-Anger (-1.0), num sents per turn (-1.1), LIWC-Unique (-1.3), *though* ratio (-1.4)
M	Clint Eastwood (*Gran Torino*)	Walt	LIWC-WC (5.4), num of sents (4.7), LIWC-Nonfl (2.1), LIWC-Incl (1.3), num of sents per turn (1.2), LIWC-Preps (1.2), *around* ratio (1.1), Reject first ratio (1.1), LIWC-Pronoun (-1.1), LIWC-Self (-1.2), *though* ratio (-1.4), concession polarity (-1.6), LIWC-Unique (-2.0)

.2 Stories

Garden (Original)

Today when I arrived at my community garden plot, it actually looked like a garden. Not a weedy mess with maybe some stuff growing in it if you know where to look. We had hit the typical mid-summer mess of fast-growing weeds and no time to do anything about it. Plus all the rain had made a huge swamp and it was hard to get a moment to work when it wasn't actively pouring. I put in a bunch of time this past week, and it's paying off. Along with free-standing non-weed-choked plants, I have now re-planted three of the beds with salad greens, spinach, and chard. And while the viability of the seeds was questionable, I accidentally unearthed some from the bed I planted 2 days ago and they had already started to sprout! This marks the first time I have reclaimed the garden from a weed problem and turned it back into a productive garden. Other years I've never managed to get the late summer planting done. I would've liked to get salad greens in 3 weeks ago, to harvest baby greens for salads along the way, but I'm still pretty pleased. I've also got a few ideas for improving the garden next year. For one thing, the radishes fall down all over the place and make a mess when they go to seed, and they really could be removed once flea-beetle season has passed, which I didn't think to do this year. **They're attracting cool butterflies**, but I might be able to do that with some less floppy flowers.

Scheherazade & EST	**PyPer: Monologue to Dialogue (M2D)**	**PyPer: M2D + Stylistic Parameters**
The radishes charmed the butterflies. The communal garden was weedy. Rained. The communal garden was swampy. Rained. The productive narrator planted the plants. The narrator planted the chards the lettuces and the spinach. The pleased narrator did not expect for the chards the lettuces and the spinach to grow. The chards the lettuces and the spinach sprouted. The narrator mistakenly dug the chards the lettuces and the spinach. The surprised narrator saw for the chards the lettuces and the spinach to sprout. The communal garden was not weedy. The communal garden was not swampy. The communal garden was productive. The communal garden was productive. The narrator was proud. The eager narrator wanted to reap the lettuces. The radishes were droopy. The narrator planned to remove the radishes. The thoughtful narrator thought the flowers charmed the butterflies.	**Speaker 1:** The radishes charmed the butterflies. The communal garden was weedy. Rained. The communal garden was swampy. **Speaker 2:** Rained. The productive narrator planted the plants. The narrator planted the chards the lettuces and the spinach. The pleased narrator did not expect for the chards the lettuces and the spinach to grow. **Speaker 1:** The chards the lettuces and the spinach sprouted. The narrator mistakenly dug the chards the lettuces and the spinach. **Speaker 2:** The surprised narrator saw for the chards the lettuces and the spinach to sprout. The communal garden was not weedy. **Speaker 1:** The communal garden was not swampy. The communal garden was productive. The communal garden was productive. **Speaker 2:** The narrator was proud and wanted to reap the lettuces. The radishes were droopy. The narrator planned to remove the radishes. **Speaker 1:** The thoughtful narrator thought the flowers charmed the butterflies.	**Speaker 1:** The radishes charmed them. The communal garden was weedy. Rained. The communal garden was swampy. **Speaker 2:** Rained, very well. She planted the well, plants. Typical. Yeah, she planted the chards the lettuces and the spinach , right! Ok, she did not expect for the chards the lettuces and the spinach to grow. **Speaker 1:** The chards the lettuces and the spinach sprouted. She mistakenly dug the chards the lettuces and the spinach. **Speaker 2:** Oh she saw I am delighted to say that for the chards the lettuces and the spinach to sprout. Typical. The communal garden was not not weedless. **Speaker 1:** The communal garden was not swampy. The communal garden was productive. The communal garden was productive. **Speaker 2:** Yeah, who was proud, very well and wanted to reap the lettuces? Oh you might be interested in knowing that the radishes were droopy let's see, that ...„ you know. I mean, she planned to remove the radishes. **Speaker 1:** She thought the flowers charmed them.

Figure .1: The Garden Story Example Highlighting Some Differences

Protest (Original)

The protesters apparently started their protest at the Capitol Building then moved to downtown. We happened to be standing at the corner of 16th and Stout when somebody said that the Police were getting ready to tear-gas a group of demonstrators. We looked around the corner and there were Police everywhere. They had blockaded the whole street, and shut down the light rail. It turned out there were about 200 protesters who were demonstrating in the streets. Supposedly, the goal of the protesters was to block streets and traffic. The protesters were demonstrating primarily against the Iraq War and more generally against corporate power, which, in their minds, inflicts both political parties. The Police got after them to move, and about half of them went into a parking garage. The Police sealed the garage and went in after them. It was shocking to see the number of Police involved. They easily out-numbered the protesters by about 10-1. The protesters were mostly in their late teens and early twenties. The Police contingent included SWAT teams, riot teams, and Police on horses. The organization of security is different than in Boston. In Boston, there were many law enforcement personnel throughout the convention area, but they tended to be more spread out. When we walked out of our hotel this afternoon, the first thing we saw was a SWAT team outside the hotel, in full riot control gear. We saw several such teams in the two block walk to the 16th Street Mall. There were vehicles loaded with about 10-12 police patrolling through the streets–entire groups of Police nearly everywhere you looked. Frankly, it seemed a bit of an over-reaction. On the other hand, it's clear that any attempt to block traffic will not be allowed, which is quite understandable. I didn't see the Police do anything really out of line–in fact, they seemed thoroughly professional. Still, it was a bit abrupt to see three companies of Police actually marching through the streets of Denver.

Scheherazade & EST	**PyPer: Monologue to Dialogue (M2D)**	**PyPer: M2D + Stylistic Parameters**
The protesters started to protest at the Capitol Building. The protesters moved toward the downtown. Narrator stood at Stout and 16th. The informed person said the police used the tear gas on the protesters. Narrator saw the police. The protesters protested in the street in order for the protesters to block the street. The police obstructed the street. The police tried to move the protesters. The protesters moved into the parking lot. The police blocked the parking lot. The police pursued the protesters. Concerned narrator returned to the hotel. Adventurous narrator left the hotel. Narrator saw the armed SWAT team. Curious narrator walked toward the mall. Narrator saw the armed SWAT team. Narrator saw the police cars. Narrator thought the police was not oppressive.	**Speaker 1:** The protesters started to protest at the Capitol Building. The protesters moved toward the downtown. **Speaker 2:** Narrator stood at Stout and 16th. The informed person said the police used the tear gas on the protesters. **Speaker 1:** Narrator saw the police. The protesters protested in the street in order for the protesters to block the street. **Speaker 2:** The police obstructed the street and tried to move the protesters. **Speaker 1:** The police pursued the protesters, blocked the parking lot and pursued the protesters. **Speaker 2:** The protesters moved into the parking lot. **Speaker 1:** Concerned narrator returned to the hotel. **Speaker 2:** Adventurous narrator left the hotel. Narrator saw the armed SWAT team. **Speaker 1:** Narrator saw the police cars and saw the armed SWAT team. **Speaker 2:** Curious narrator walked toward the mall. **Speaker 1:** Narrator thought the police was not oppressive.	**Speaker 1:** The protesters started to protest at the Capitol Building. The protesters moved toward the downtown. **Speaker 2:** Unfortunately, she stood at corner, especially actually, as it were. Actually, the quite informed person said it used the tear gas on the protesters. **Speaker 1:** She saw it. The protesters protested in the street in order for the protesters to barricade the street. **Speaker 2:** I do not know but, it obstructed the street and tried to move the protesters. **Speaker 1:** It pursued the protesters, blocked the parking lot and pursued the protesters. **Speaker 2:** It seems to me that what moved into the parking lot? Typical. **Speaker 1:** The demonstrators moved into the parking lot. She returned to the hotel. **Speaker 2:** Actually, she left the hotel. I might be wrong but, she saw the largely armed SWAT team, as it were. **Speaker 1:** She saw the police cars and saw the armed SWAT team. **Speaker 2:** She walked toward the mall, as it were. Typical. What happenned next? **Speaker 1:** She thought it was not oppressive.

Figure .2: The Protest Story Example Highlighting Some Differences

Squirrel (Original)

This is one of those times I wish I had a digital camera. We keep a large stainless steel bowl of water outside on the back deck for Benjamin to drink out of when he's playing outside. His bowl has become a very popular site. Throughout the day, many birds drink out of it and bathe in it. The birds literally line up on the railing and wait their turn. Squirrels also come to drink out of it. The craziest squirrel just came by- he was literally jumping in fright at what I believe was his own reflection in the bowl. He was startled so much at one point that he leap in the air and fell off the deck. But not quite, I saw his one little paw hanging on! After a moment or two his paw slipped and he tumbled down a few feet. But oh, if you could have seen the look on his startled face and how he jumped back each time he caught his reflection in the bowl!

Scheherazade & EST	**PyPer: Monologue to Dialogue (M2D)**	**PyPer: M2D + Stylistic Parameters**
I gently placed the steely bowl on the deck in order for Benjamin to drink the bowl's water. The steely bowl was popular. The colorful birds drank the bowl's water. The birds playfully bathed themselves in the steely bowl. The colorful birds organized themselves on the deck's railing in order for the birds to wait. The frisky squirrels drank the bowl's water. The squirrel cautiously approached the steely bowl. The crazy squirrel was startled because the squirrel saw the squirrel's reflection. The crazy squirrel leaped because the squirrel was startled. The crazy squirrel fell over the deck's railing because the squirrel leaped because the squirrel was startled. The crazy squirrel desperately held the deck's railing with the squirrel's paw. The squirrel's paw unfortunately slipped off the deck's railing. The crazy squirrel ultimately fell.	**Speaker 1:** I gently placed the steely bowl on the deck in order for Benjamin to drink the bowl's water. **Speaker 2:** The steely bowl was popular. The colorful birds drank the bowl's water. The birds playfully bathed themselves in the steely bowl. **Speaker 1:** The colorful birds organized themselves on the deck's railing in order for the birds to wait. The frisky squirrels drank the bowl's water. **Speaker 2:** The squirrel cautiously approached the steely bowl. The crazy squirrel was startled because the squirrel saw the squirrel's reflection. **Speaker 1:** The crazy squirrel leaped because the squirrel was startled. The crazy squirrel fell over the deck's railing because the squirrel leaped because the squirrel was startled. **Speaker 2:** The crazy squirrel desperately held the deck's railing with the squirrel's paw. The squirrel's paw unfortunately slipped off the deck's railing. **Speaker 1:** The crazy squirrel ultimately fell.	**Speaker 1:** She gently placed the steely bowl on the deck in order for Benjamin to drink the bowl's water. **Speaker 2:** Right, they drank the bowl's water , somewhat you are kidding, right? Ok, what playfully bathed themselves in the very steely bowl? I see, what was pretty popular? **Speaker 1:** They organized themselves on the deck's railing in order for them to wait. They drank the bowl's water. **Speaker 2:** It seems that it cautiously approached the steely bowl. It was well, startled because it saw it's reflection. **Speaker 1:** It leaped because it was startled. It fell over the deck's railing because it leaped because it was startled. **Speaker 2:** It desperately held the deck's railing with it's paw! Basically, it's paw unfortunately literally slipped off the deck's railing. **Speaker 1:** It ultimately fell.

Figure .3: The Squirrel Story Example Highlighting Some Differences

Bug Out for Blood (Original)

Bug out for blood the other night, I left the patio door open just long enough to let in a dozen bugs of various size. I didn't notice them until the middle of the night, when I saw them clinging to the ceiling. Since I'm such a bugaphobe, I grabbed the closest object within reach, and with a rolled-up comic book I smote mine enemies and smeared their greasy bug guts. All except for the biggest one. I don't know what it was; it was one of those things you see skimming the surfaces of lakes, with a legspan of a few inches. I only clipped that one, taking off one of its limbs. But it got away before I could finish the job. So now there's a five-limbed insect lurking in the apartment, no doubt looking for some vengeance against me. I'm looking around corners, checking the toilet before sitting down, checking the bowl before taking another scoop of cereal, wondering when it's going to jump out. All this, and the cicadas haven't even arrived yet.

Scheherazade & EST	**PyPer: Monologue to Dialogue (M2D)**	**PyPer: M2D + Stylistic Parameters**
I recently momentarily openned my patio's door. The slimy bugs quietly entered my apartment. I did not initially notice that the slimy bugs quietly entered my apartment. I peacefully slept. I overnight awoke. I shockingly saw the slimy bugs was on my apartment's ceiling. I grabed the reachable thing because the bugs scared me. The slimy bugs were my enemy. I smeared the greasy bugs's innards. I managed to kill every bug. I grazed the bugs's leader with the rolled comic book. I removed the tiny bugs's leader's limb. I angrily hit the bugs with the rolled comic book. The big bugs's leader now lurked in my apartment. The bugs's leader was the three-legged insect. The big bugs's leader escaped me. The big bugs's leader undoubtedly wanted to retaliate against me. I looked around every corner of my apartment. I checked my toilet seat for the big bugs's leader in order for me to sit down on my toilet seat. I in due course expected for the big bugs's leader to jump toward me. The cicadas did not even yet arrive to my apartment. Every action of my story notably happenned.	**Speaker 1:** I recently momentarily openned my patio's door. The slimy bugs quietly entered my apartment. **Speaker 2:** I overnight awoke, peacefully slept and overnight awoke. **Speaker 1:** I did not initially notice that the slimy bugs quietly entered my apartment. **Speaker 2:** I shockingly saw the slimy bugs was on my apartment's ceiling. I grabed the reachable thing because the bugs scared me. **Speaker 1:** The slimy bugs were my enemy. I smeared the greasy bugs's innards. I managed to kill every bug. I grazed the bugs's leader with the rolled comic book. **Speaker 2:** I removed the tiny bugs's leader's limb. I angrily hit the bugs with the rolled comic book. The big bugs's leader now lurked in my apartment. **Speaker 1:** The bugs's leader was the three-legged insect. The big bugs's leader escaped me. The big bugs's leader undoubtedly wanted to retaliate against me. I looked around every corner of my apartment. **Speaker 2:** I checked my toilet seat for the big bugs's leader in order for me to sit down on my toilet seat. I in due course expected for the big bugs's leader to jump toward me. **Speaker 1:** The cicadas did not even yet arrive to my apartment. Every action of my story notably happenned.	**Speaker 1:** She recently momentarily openned she patio's door. They quietly entered she apartment. **Speaker 2:** Oh gosh she overnight awoke everybody knows that, peacefully slept and overnight awoke. **Speaker 1:** She did not initially notice that they quietly entered she apartment. **Speaker 2:** She shockingly saw them was on she apartment's ceiling. She grabed the reachable obviously, thing because they scared her. **Speaker 1:** They were she enemy. She sullied they's viscera. She managed to kill every it. She grazed they's leader with the rolled comic book. **Speaker 2:** She removed they's leader's limb. She angrily hit them everybody knows that with the rolled comic book. They's leader now lurked in she apartment. **Speaker 1:** They's leader was it. They's leader escaped her. They's leader undoubtedly wanted to retaliate against her. She looked around every corner of she apartment. **Speaker 2:** She checked she toilet seat for they's leader in order for her to sit down on she toilet seat. She in due course expected for they's leader to jump toward her. **Speaker 1:** They did not even yet arrive to she apartment. Every action of she story notably happenned.

Figure .4: The Bug Out for Blood Story Example Highlighting Some Differences

Employer Botches Training (Original)

I hope i'm not turning in to one of those people who are unemployed in the US. PF Chang's really messed up my training. It was one person really. If you would've seen this schedule i got you would understand. some of you did see it so you know what i mean. I went in last Wednesday to take what i thought was my final training class. I was told i missed it and it was the day before. I was really confused because my schedule said i was off that day so i pulled it out and showed them. Then she says thats not your schedule. I was like its not what do you mean? She said i wrote you a new one and gave it to someone to give to you. I said i never got it. who did you give it too? She didn't know. She said it was posted online also on the schedule website. I said no it wasn't cause i was just on that this morning. She looked and sure enough i was right. She said well now i need to talk to my boss and see if there are any training hours left for you. I said there has to be i didn't use them. She wouldn't even let me follow someone around again that night. So i had to go home with the fact that she said she would be calling me by the end of the week. Did she call?? NO! SO i will call on Monday morning. I don't know if i even want to go back now. She is the server manager and i'd have to deal with her all the time then. So this weekend i proceeded to apply for some jobs online. Hopefully tomorrow i will hear back from someone. This Sunday jim and i leave for Disney. I could wait to apply for more until come back from that. When i come back i have lots of stuff at tdt going on. theres some income. Does anyone know of any places who are hiring? part time or full it doesn't matter.

Scheherazade & EST	**PyPer: Monologue to Dialogue (M2D)**	**PyPer: M2D + Stylistic Parameters**
I excitedly entered PF Changs in order for the manager to train me. The manager lazily said the manager yesterday scheduled me in order for the manager to train me and I did not show up. I was confused because Schedule demonstrated my punctuality. I promptly showed Schedule to the manager. The disgruntled manager lazily said Schedule was erroneous. I insistently questioned the manager because the manager lazily said Schedule was erroneous. The manager melodramatically said the manager created New schedule and the manager gave New schedule to the employee in order for the employee to give New schedule to me. I said I did not receive New schedule. I cautiously questioned the manager about the employee's identity. The disgruntled manager did not flippantly know the employee's identity. The disgruntled manager said the manager put New schedule on the website. I saw New schedule was not on the website I quickly said New schedule was not on the website and I checked the website for New schedule. The disgruntled manager checked the website for New schedule. The disgruntled manager surprisedly saw New schedule was not on the website. The manager said the manager needed to talk to the manager's boss in order for the manager to train me. I frenziedly said the manager needed to train me. The manager did not train me. The manager said the manager called me. I disappointedly returned to the home. The manager did not unfortunately call me. I planned to call the manager. I did not however want to return to PF Changs because I disliked the manager. I ultimately searched the internet for the employment.	**Speaker 1:** I excitedly entered PF Changs in order for the manager to train me. The manager lazily said the manager yesterday scheduled me in order for the manager to train me and I did not show up. **Speaker 2:** I was confused because Schedule demonstrated my punctuality. I promptly showed Schedule to the manager. The disgruntled manager lazily said Schedule was erroneous. I insistently questioned the manager because the manager lazily said Schedule was erroneous. **Speaker 1:** The manager melodramatically said the manager created New schedule and the manager gave New schedule to the employee in order for the employee to give New schedule to me. I said I did not receive New schedule. I cautiously questioned the manager about the employee's identity. **Speaker 2:** The disgruntled manager did not flippantly know the employee's identity. The disgruntled manager said the manager put New schedule on the website. I saw New schedule was not on the website I quickly said New schedule was not on the website and I checked the website for New schedule. **Speaker 1:** The disgruntled manager checked the website for New schedule. The disgruntled manager surprisedly saw New schedule was not on the website. The manager said the manager needed to talk to the manager's boss in order for the manager to train me. **Speaker 2:** I frenziedly said the manager needed to train me. The manager did not train me and said the manager called me. I disappointedly returned to the home. The manager did not unfortunately call me. **Speaker 1:** I planned to call the manager. I did not however want to return to PF Changs because I disliked the manager. I ultimately searched the internet for the employment.	**Speaker 1:** She excitedly entered PF Changs in order for the manager to train her. She lazily said the manager yesterday scheduled her in order for the manager to train her and she did not show up. **Speaker 2:** I see, she was confused because Schedule demonstrated she punctuality, technically. She promptly literally showed Schedule to the manager. The very disgruntled manager lazily said Schedule was well, erroneous. She insistently questioned the manager because the manager lazily said Schedule was rather erroneous, you know. **Speaker 1:** The manager melodramatically said the manager created New schedule and the manager gave New schedule to the employee in order for the employee to give New schedule to her. She said she did not receive New schedule. She cautiously questioned the manager about the employee's identity. **Speaker 2:** Who did not flippantly know the employee's identity? I thinks the really disgruntled manager said the manager put New schedule on the website. She saw New schedule was not on the website I mean, she quickly said New schedule was not on the website and she checked the website for New schedule. **Speaker 1:** The disgruntled manager checked the website for New schedule. The disgruntled manager surprisedly saw New schedule was not on the website. The manager said the manager needed to talk to the manager's boss in order for the manager to train her. **Speaker 2:** I that thinks she frenziedly said the manager needed to train her. The manager did not train her and said the manager called her. She disappointedly returned to the home. The manager did not unfortunately call her. **Speaker 1:** She planned to call the manager. She did not however want to return to PF Changs because she disliked the manager. She ultimately searched the internet for the employment.

Figure .5: The Employer Botches Training Story Example Highlighting Some Differences

Storm (Original)

That was one hell of a storm, the biggest to hit Baton Rouge. The entire city was out of power the first few days, and it took seven days for power to be restored in my neighborhood.? The damage was widespread across Baton Rouge, the wind had mangled store fronts and signs, and knocked over trees crushing houses and damaging power lines.? A curfew has been in place most of the week upon threat of arrest. Blackhawk, Chinook, news, police, and Coast Guard helicopters could be seen or heard in the skies hourly. The National Guard came in huge convoys on the interstate and set up distribution centers for ice, tarps, and MRE's.? Luckily we had prepared with enough food to last us the week, and enough gas to run the generator for a few days.? It took days more for the gas stations and stores to start opening again, and several-hour-long lines would form outside.? We ended up driving an hour out of town just to restock on supplies.Things are almost back to normal now. Only about 30-40% of the city is left without power and most of the stores and gas stations are back online.? Through TV and radio we were kept up to date on local news, but as far as national news I've lived in a virtual black hole. My power was restored last night and it's going to take awhile to get things rolling again. School has started today, and they want to take away some of our Saturdays for make-up classes. And if that weren't bad enough, hurricane Ike is on the horizon and could just as well hit us again, causing more power and scheduling problems. Keep your eyes peeled the next couple of days for some photos I took of the damage.

Scheherazade & EST	**PyPer: Monologue to Dialogue (M2D)**	**PyPer: M2D + Stylistic Parameters**
The big storm hit in the Baton Rouge. The city did not have the electricity. The city restored the electricity in the neighborhood. The damage was widespread in the Baton Rouge. The wind mangled the storefront. The wind mangled the sign. The wind bumped the tree. The tree crushed the house. The tree damaged the power line. The police set the rule in order for the police to threaten the people against the arrest. I heard the police. I heard the US Coast Guard's heliport in the sky. The National Guard came in the huge vehicle on the interstate. The National Guard set the center. The National Guard distributed the ice. The National Guard distributed the tarp. The National Guard distributed the meal's package. I prepared the food. I stored gas enough for the generator to be last. The gas station openned. The store openned. I drove out of the town in order for me to buy the thing. I heard about the local news from the TV from the radio. I did not hear the national news. The electric company restored the electricity. The school started. The academic administrator wanted to take the class. I took the picture about the damage.	**Speaker 1:** The big storm hit in the Baton Rouge. The city did not have the electricity and restored the electricity in the neighborhood. **Speaker 2:** The wind mangled the sign, mangled the storefront and mangled the sign. **Speaker 1:** The damage was widespread in the Baton Rouge. **Speaker 2:** The tree damaged the power line, crushed the house and damaged the power line. The wind bumped the tree. The police set the rule in order for the police to threaten the people against the arrest. **Speaker 1:** I heard the police and heard the US Coast Guard's heliport in the sky. The National Guard came in the huge vehicle on the interstate. **Speaker 2:** The National Guard set the center and distributed the ice. The National Guard distributed the ice and distributed the meal's package. The National Guard distributed the tarp. I prepared the food. **Speaker 1:** I stored gas enough for the generator to be last. The gas station openned. The store openned. I drove out of the town in order for me to buy the thing. **Speaker 2:** I heard about the local news from the TV from the radio. I did not hear the national news. The electric company restored the electricity. **Speaker 1:** The school started. The academic administrator wanted to take the class. I took the picture about the damage.	**Speaker 1:** The big storm hit in the Baton Rouge. It did not have the electricity and restored the electricity in the neighborhood. **Speaker 2:** The wind mangled the sign, very well, you know, mangled what and mangled the sign. **Speaker 1:** The wind mangled the sign, mangled the storefront and mangled the sign. The damage was widespread in the Baton Rouge. **Speaker 2:** Right, the tree damaged the power line, crushed the house and damaged the power line. I might be wrong but, the wind bumped the tree. Typical. It set the well, rule in order for it to threaten it against the arrest. **Speaker 1:** It heard it and heard the US Coast Guard's heliport in the sky. The National Guard came in the huge vehicle on the interstate. **Speaker 2:** The National Guard set the center and distributed the ice. The National Guard distributed the well, ice and distributed the meal's package. Typical. Darn i see, I do not remember what happenned next, you know! **Speaker 1:** The National Guard distributed the tarp. It prepared the food. It stored gas enough for the generator to be last. The gas station openned. The store openned. It drove out of the town in order for it to buy the thing. **Speaker 2:** Darn oh it heard I am delighted to say that about the local news from the TV from the radio. It did not hear the national news. The electric company restored the electricity , oh God did not it? **Speaker 1:** The school started. The academic administrator wanted to take it. It took the picture about the damage.

Figure .6: The Storm Story Example Highlighting Some Differences

.3 Mapping to PyPer

Table .2: Mapping: LIWC Categories Examples

PyPer Param	LIWC category	PyPer Param	LIWC category
near-expletives	liwc-swear liwc-anger	low-expletives	liwc-swear liwc-anger
emph-actually	liwc-certain	emph-exclamation	liwc-excl
emph-really	liwc-certain	emph-great	liwc-assent
emph-you-know	liwc-filler	emph-particularly	liwc-certain
emph-technically	liwc-certain	emph-literally	liwc-certain
emph-quintessential	liwc-certain	emph-essentially	liwc-certain liwc-i
emph-somewhat	liwc-tentat	emph-very	liwc-certain
emph-especially	liwc-certain	emph-roughly	liwc-tentat
in-group-marker	liwc-family,liwc-friends, liwc-we, liwc-incl	init-reject	liwc-tentat
competence-mitigation	liwc-self,liwc-negemo, liwc-sad	ack-i-see	liwc-filler, liwc-tentat, liwc-filler
ack-well	liwc-filler, liwc-tentat, liwc-filler	ack-yeah	liwc-assent, liwc-filler
ack-right	liwc-assent, liwc-filler	ack-oh	liwc-nonfl
ack-ok	liwc-assent, liwc-filler	soften-adjective	liwc-tent
down-kind-of	liwc-tentat	down-sort-of	liwc-tentat
down-unfortunately	liwc-negemo, liwc-sad	down-might-be-interested	liwc-tentat
down-like	liwc-tentat,liwc-filler	down-i-mean	liwc-filler
down-somewhat	liwc-tentat	down-mmhm	liwc-nonfl, liwc-filler
down-around	liwc-tentat	down-i-think	liwc-tentat, liwc-filler
down-subord	liwc-tentat, liwc-filler	verbosify	liwc-wps, liwc-wc, liwc-sixltr, liwc-article
simplify	liwc-wps, liwc-wc	actor-pronouns	liwc-pronoun, liwc-ppron, liwc-ipron, liwc-they, liwc-shehe
ask-question	liwc-qmarks,liwc-anx,liwc-tentat,liwc-discrep	ask-n-answer	liwc-qmarks, liwc-wps
provoke-question	liwc-qmarks, liwc-wps, liwc-wc	request-confirmation	liwc-qmarks
negate-polarity	liwc-negate	down-i-mean	liwc-feel
down-i-think	liwc-feel	init-reject	liwc-feel
soften-adjective	liwc-feel	down-might-be-interested	liwc-posemo
ack-yeah	liwc-posemo	ack-ok	liwc-posemo
ack-very-well	liwc-posemo	emph-delighted	liwc-posemo
emph-great	liwc-posemo	soften-adjective	liwc-posemo
down-i-think	liwc-insight		

Table .3: Mapping: Dialogue Act Categories Examples

PyPer Param	Features	PyPer Param	Features
tag-question	dialact-ynQuestion-ratio, dialact-last-ynQuestion-ratio, dialact-first-ynQuestion-ratio	ask-question	dialact-whQuestion-ratio, dialact-last-whQuestion-ratio, dialact-first-whQuestion-ratio, dialact-Clarify-ratio, dialact-last-Clarify-ratio, dialact-first-Clarify-ratio
init-reject	dialact-first-Reject-ratio, dialact-last-Reject-ratio, dialact-Reject-ratio	competence-mitigation	dialact-first-Reject-ratio
ack-yeah	dialact-last-yAnswer-ratio, dialact-last-Accept-ratio, dialact-first-Accept-ratio, dialact-Accept-ratio	ack-ok	dialact-last-yAnswer-ratio, dialact-last-Accept-ratio, dialact-first-Accept-ratio, dialact-Accept-ratio
ask-n-answer	dialact-last-nAnswer-ratio, dialact-first-nAnswer-ratio, dialact-nAnswer-ratio, dialact-first-yAnswer-ratio, dialact-last-yAnswer-ratio, dialact-yAnswer-ratio	down-i-mean	dialact-Emotion-ratio, dialact-first-Emotion-ratio
down-i-think	dialact-Emotion-ratio, dialact-first-Emotion-ratio	init-reject	dialact-Emotion-ratio
soften-adjective	dialact-Emotion-ratio	tag-question	dialact-last-Emotion-ratio, dialact-last-Emotion-ratio
emph-actually	dialact-Emphasis-ratio	emph-exclamation	dialact-last-Emphasis-ratio
emph-basically	dialact-Emphasis-ratio	emph-particularly	dialact-Emphasis-ratio
emph-literally	dialact-Emphasis-ratio, dialact-last-Em	emph-pretty	dialact-Emphasis-ratio
emph-essentially	dialact-Emphasis-ratio	emph-very	dialact-Emphasis-ratio
emph-especially	dialact-Emphasis-ratio		

Table .4: Mapping: Other Categories Examples

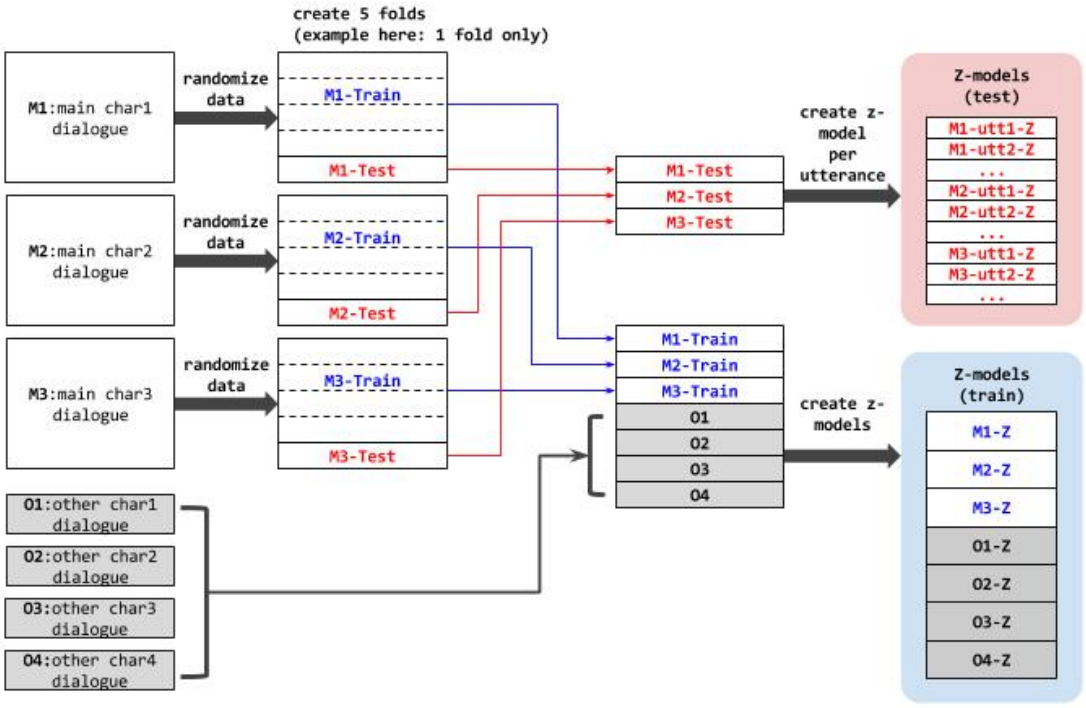
PyPer Param	Features	PyPer Param	Features
contractify	verb-verb-APOSTROPHEm-ratio, verb-verb-APOSTROPHEre-ratio	tag-question	tag-ratio
verbosify	avg-content-wlen, tokesnperutt, wordsperutt, tokenspersent, wordspersent	init-reject	polarity-percent-neg
down-unfortunately	polarity-percent-neg	down-kind-of	polarity-percent-neg
down-sort-of	polarity-percent-neg	down-subord	polarity-percent-neg
simplify	keywordsperword	paraphrase	verb-strength
ack-yeah	category-ack-ratio	ack-right	category-ack-ratio
ack-oh	category-ack-ratio	ack-ok	category-ack-ratio
ack-i-see	category-ack-ratio	ack-well	category-ack-ratio
down-i-mean	category-fill-ratio	emph-you-know	category-fill-ratio
down-somewhat	category-soft-ratio	down-quite	category-soft-ratio
down-around	category-soft-ratio	down-rather	category-soft-ratio
down-i-think	category-soft-ratio	down-subord	category-soft-ratio
down-subord	word-it-seems-ratio-allw, word-it-seems, word-it-seems-ratio-catw	emph-basically	word-basically
near-expletives	word-darn-ratio-allw, word-darn-ratio-catw, word-ass, word-bitch, word-bitch-ratio-catw, word-sucks-ratio-allw, word-piss-ratio-catw, word-piss-ratio-allw, word-sucks-ratio-catw	in-group-marker	word-pal-ratio-allw
down-sort-of	word-sort-of		

.4 TV Character Models Objective Evaluation: Cross Validation

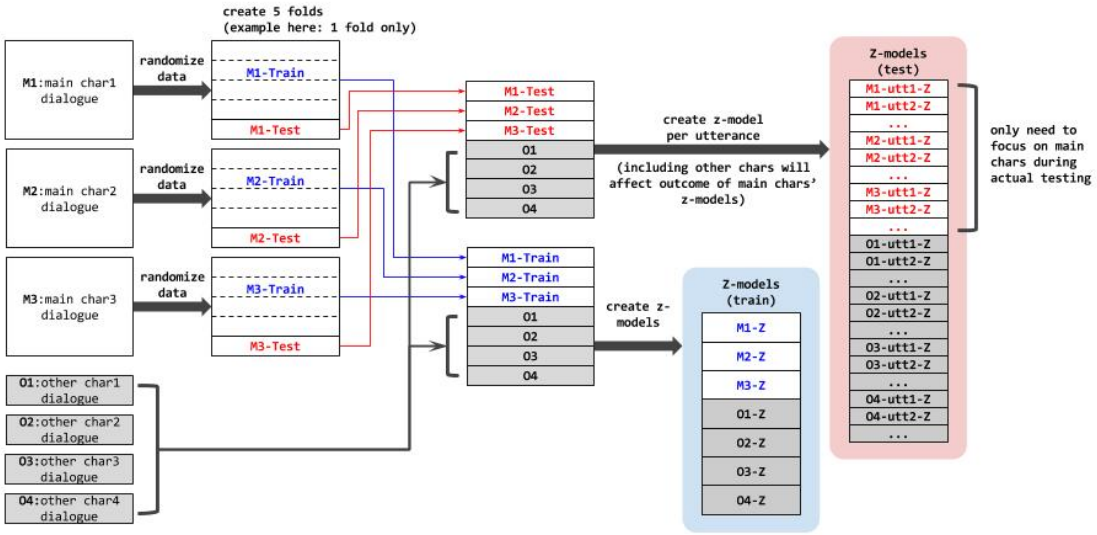
We perform the verification using 5-fold cross validation on randomized data. Figure .7 shows how the data is divided in one fold to create test/train Z-models using an example of three main characters and four non-main characters. In Figure .7a, viewing from left to right, each main character's dialogue is randomized before creating five folds. Four folds are used for training and one fold is used for testing. Combined with non-main characters' dialogue (grayed out boxes on the lower half of the figure), the trained Z-models are created for each character. This is noted on the figure inside the blue box with labels: `M1-Z, M2-Z, M3-Z, O1-Z, ..., O4-Z`. Similarly, we combined the testing folds, separated into individual utterances, then created Z-models for each utterance. This is noted on the figure inside the red box with labels: `M1-utt1-Z, M1-utt2-Z, ..., M2-utt1-Z, M2-utt2-Z`, ..., etc.

A slightly different version, shown in Figure .7b, included non-main characters' dialogue in the testing folds before separating into utterances and creating Z-models. Recall that Z-models are created using the average and standard deviation of the data. By adding non-main characters' dialogue we effectively added some noise into the test set to see how (and if) the results get affected. When performing the actual testing we only need to look at main characters' utterance Z-models. It turns out that it did not really affect the results, which will be shown in later sections.

The LM process is similar to that of Z-models, except we simply used the test set utterances as-is since the SRILM toolkit takes text as input to test the trained models. The process is shown in Figure .8. A slightly different version of LM



(a) Test set containing only main characters.



(b) Test set containing main and non-main characters.

Figure .7: Cross Validation for Z-Models

uses LIWC-tagged text. This means that each word is tagged with the LIWC categories. For example, the word "I" belongs to LIWC categories: personal

pronoun (ppron), pronoun, word count (wc), function words (funct), dictionary words (dict), and first person singular (i). Here is an example of one sentence and its LIWC-tagged version from our data:

Regular text: `I think this is the place.`

LIWC-tagged text:

I_ppron_pronoun_wc_funct_dict_i

think_insight_wc_present_verb_dict_cogmech

this_dict_pronoun_funct_ipron_wc

is_auxverb_wc_present_verb_dict_funct

the_article_dict_funct_wc

place_dict_relativ_space_wc

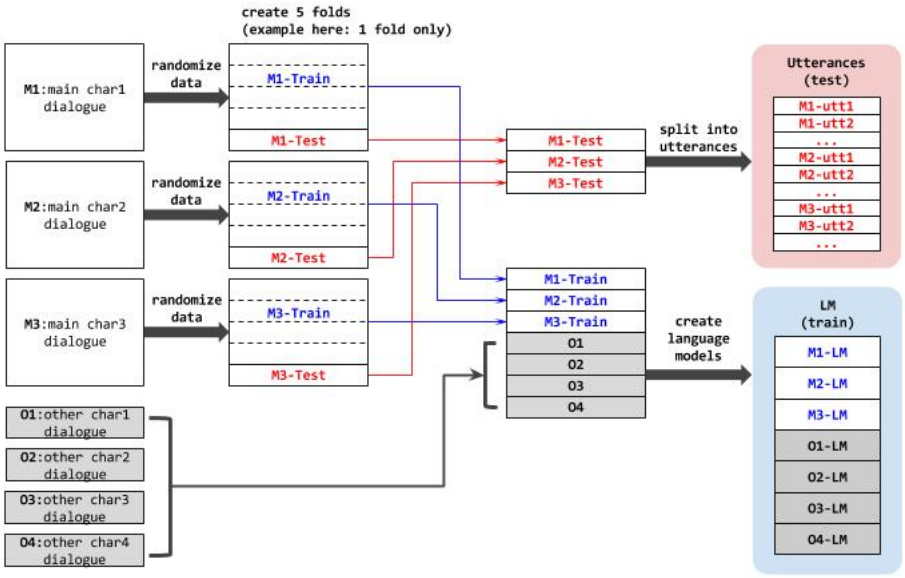


Figure .8: Cross Validation for LM

.5 Characters' Similarity Count MTurk Results

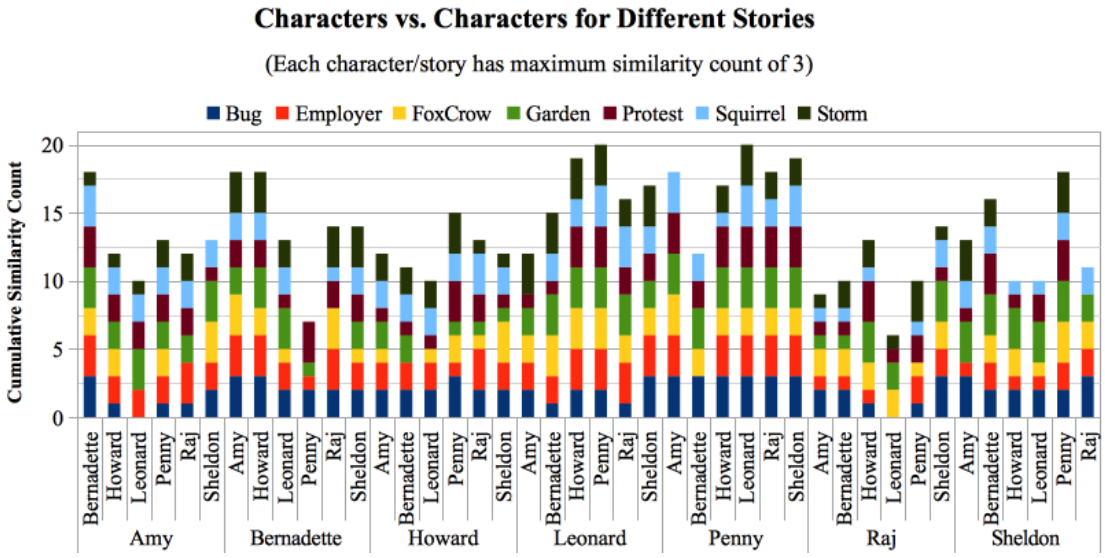


Figure .9: Characters vs. Characters for Different Stories MTurk Results

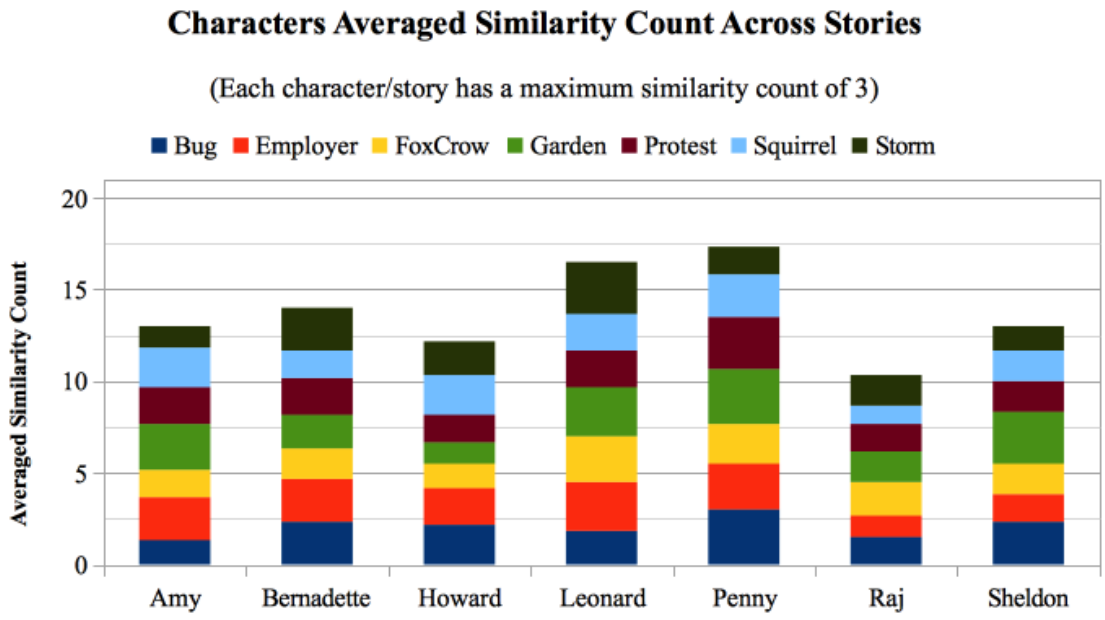


Figure .10: Characters' Averaged Similarity Count Across Stories MTurk Results

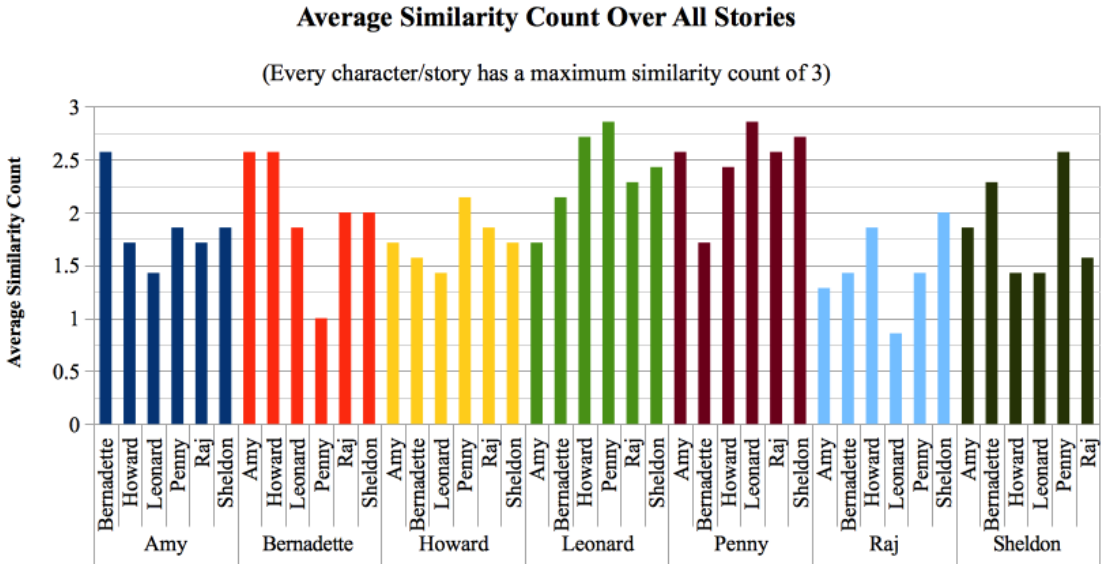


Figure .11: Average Similarity Count over All Stories MTurk Results

.6 Character Analysis

Table .5: Sheldon vs. Leonard Comments on Dialogue

yellow for Sheldon traits; green for less-Sheldon (Leonard) traits

	Sheldon vs. Leonard: Sheldon-modeled dialogue more similar to Sheldon
1	Dialogue 2 is a bit short. I could better picture Sheldon speaking longer lines, kind of ranting.
2	Seeing the words I don't know seems kind of out of character for him which would make Dialogue 1 the better choice.
3	When he says, "mmhm" it really doesn't sound like him. Also the fact that he ends every sentence short does not suite his speech pattern.
4	Dialogue 1 definitely sounds more like Sheldon's character as it sounds overly complicated and more drawn out.
5	Sheldon is verbose, analytical, precise and socially awkward. Sheldon is not inclined to say "oh," or "mmhm" or similar non-words. He is also rarely at a loss (unless agitated).
6	Sheldon seems sort of wooden or clinical in the sample dialogue. I don't think he would use casual language like "Mmhm" often. S2 seems more relaxed in Dialogue 2, and I think Sheldon would be more formal than that.
7	I picked S2 from S2 in Dialogue 1 is almost a perfect fit for Sheldon, with the precision of wording, and use of elaborate vocabulary. The second line in this dialogue I could actually picture Sheldon saying perfectly.
8	"As it were" is a phrase Sheldon would use.
9	The speaker in Dialogue 2 seems too confused to be Sheldon.
	Sheldon vs. Leonard: Leonard-modeled dialogue more similar to Sheldon
1	Sheldon states things pretty clearly in the sample dialogue. I feel like S2 in Dialogue 2 is more matter-of-fact, and Sheldon's character would probably be straightforward like that.
2	I can picture Sheldon saying "Mmhm" arrogantly in the last line of Dialogue 2. This dialogue seems more like Sheldon as it is clear and unhesitant.
3	I think the shorter, more direct sentences in Dialogue 2 are more like Sheldon. He tends to be more to the point with his speech patterns. I also feel like he would use words like "technically" as S2 in dialogue 2 does.
4	Sheldon seems like he can be scornful of people who are less informed or intelligent. The line, "The cheese fell I thought everybody knew that," seems exactly like something Sheldon would say.
5	The to-the-point sentences in dialogue 2 seem more like Sheldon, and I feel like the other mannerisms (implying that certain things are obvious) go well with his character.
6	This one really does sound like him, the wording and the expressions used fits for him perfectly, compared to one which is somewhat styled the way he would talk, but 2 just fits better.
7	The second line of S2 in Dialogue 2 sounds exactly like Sheldon. He is arrogant and ridiculously smart, and he wouldn't miss an opportunity to make others feel less intelligent than him.
8	This one was a little more challenging for me. Generally, it doesn't seem like Sheldon uses many long and complex sentences in the example dialogue, which makes me initially lean towards Dialogue 2 due to its brevity and straightforwardness. However, there are some mannerisms there (like the "oh" and "come on") that I'm not sure Sheldon would say. Overall, I think the long strings of adjectives are least reflective of his character as presented in the examples, so I chose Dialogue 2.
9	Sheldon is verbose, analytical, precise and socially awkward. Both dialogues introduce rhetorical flourishes ("as it were" in 1 and "ok," and "I might be wrong" in 2) that don't feel true to Sheldon. The precision of 2, however, has a Sheldon-like feel to it.
10	I could see both dialogues fitting for his speech pattern, but 2 is more consistent and straight forward, also with questioning and then answering is sounds more like him.
11	Speaker 1 corrects mundane details the other speaker got wrong in Dialogue 2, about the National Guard. This fits with Sheldon's nitpicking and attention to detail.

Table .6: Sheldon (in comparison to Penny) Full Worker Comments on Dialogue

yellow for Sheldon traits; green for less-Sheldon (Penny) traits

	Sheldon vs. Penny: Sheldon-modeled dialogue more similar to Sheldon
1	Dialogue 2 is way to free flowing and easy going compared to his usually speech pattern, making 1 the better fit.
2	S2 uses a very conversational tone in Dialogue 2 – "Mmhm" and "Darn ok," for example. Sheldon seems like he doesn't always understand social conventions, like in the sample when he doesn't get why people would "chat."
3	Dialogue 2 has to many uses of the word damn and darn to really fit with how he talks. He talks more with larger, complicated words.
4	I don't think that Sheldon would use some of the language that is in dialogue 2 such as "damn" and "oh god." I also don't think that the questions in the middle of some of the speech block match his patterns in the example dialogue. Because of this, I think dialogue 1 is more representative.
5	Once again I do not see him saying "Damn," or "mhmm," the whole last part of S2 doesn't suite his speech pattern.
6	The language in Dialogue 2 is very informal, and I don't think Sheldon would speak that way. I can't imagine him ever saying, "Damn mmhm."
7	Although dialogue 1 has some traits that don't seem to fit Sheldon, like the more complex sentence structures, I think it is more in line with his character overall than the overly casual and more "stream of consciousness" style of dialogue 2. I can't see his character starting a sentence with "damn mmmhm" though I could more easily see him using "as it were"
8	Sheldon is verbose, analytical, precise and socially awkward. D2 includes a lot of emotion-based statements (great, damn, exclamatory statements) that wouldn't be appropriate for Sheldon.
9	S2 in Dialogue 2 is very enthusiastic. I don't think Sheldon is very connected to emotions, so he probably wouldn't say things like "Oh god" or use even mild swearwords.
10	I think that Dialogue 2 sounds a bit too animated and relaxed to be Sheldon. He is typically straightforward without a lot of fluff in his lines.
11	Dialogue has to many points where he uses simple words and add on conjunction words that are not needed so dialogue 1 would be best.
12	S2 in Dialogue 2 just doesn't have the decisive and confident tone that Sheldon typically portrays. Dialogue 1 is very fitting to Sheldon.
13	Sheldon would not say "yeah" or "you know", as illustrated in Dialogue 2.
14	Dialogue 2 seems to be asking for permission to have an opinion too much.
15	Both dialogues have sentences that I think are a little long for Sheldon's speech patterns, but dialogue 1 does this with a lot of adjectives while dialogue 2 does it with intejectory phrases ("very well", "you know," "I see," etc.). I think the former is more consistent with Sheldon in the examples than the latter is. He tends to express his thoughts pretty completely without a lot of interruption.
16	Sheldon is verbose, analytical, precise and socially awkward. Dialogue 1's S2 is far more precise in his language than Dialogue 2. Dialogue 2 also uses colloquialisms, mild swearing (damn) and other embellishments that aren't part of Sheldon's way of speaking.
17	The cussing in Dialogue 2 automatically makes it not his speech pattern.
18	Use of the word "damn" in Dialogue 2 is a red flag, as Sheldon rarely swears. While it may slightly fit with Sheldon's Texas roots, this is something he tries to hide. Because of that, I went with Dialogue 1.
	Sheldon vs. Penny: Penny-modeled dialogue more similar to Sheldon
1	"mmhm..." I can picture coming from Sheldon in an irritated manner. "...you are kidding, right" would be said by Sheldon in an arrogant and condescending manner. Dialogue 2 is the better of these two.
2	S2's first line in Dialogue 2 sounds just like Sheldon. "You might be interested in knowing..." sounds like an arrogant Sheldon line, followed by the "Oh God..." I can actually picture Sheldon saying this line.
3	"You might be interested in knowing..." is used twice in Dialogue 2, and would be something Sheldon might say to make another person feel inferior.

Table .7: Sheldon (in comparison to Raj) Full Worker Comments on Dialogue

yellow for Sheldon traits; green for less-Sheldon (Raj) traits

	Sheldon vs. Raj: Sheldon-modeled dialogue more similar to Sheldon
1	Dialogue 1 is very complex and matter of fact, sounding much like the intelligent and arrogant Sheldon.
2	Although Dialogue 2 is very plain and straightforward which would usually suite well for him, the part where he says Quiet come on, does not.
3	I can imagine Sheldon might say, "I thought everybody knew that," like S2 in Dialogue 2, but I can't imagine him saying "Oh gosh," and the language in Dialogue 2 is more simple in general. Sheldon often uses advanced vocabulary, so I'd guess he'd speak more like S2 in Dialogue 1: "quintessential," "particularly," etc.
4	The use of cuss words and the use of oh darn does not really sound like him. Also the words flippantly doesn't sound like him, and the general style of speech does not fit with him.
5	I don't feel like Sheldon would use the phrases that are in dialogue 2 such as "oh god," "oh gosh," "damn," and "darn." At the least, he wouldn't use them so often. Even though I think he uses shorter sentences more like dialogue 2, I think overall Dialogue 1 is more like him because of the word choices.
6	I wouldn't pick 2 because the dialogue is very simple as well as the word Damn being involved. It just doesn't suite the pattern he has.
7	Sheldon doesn't seem like the type of character who would say "damn" frequently like S2 does in Dialogue 2. If it weren't for that, though, I probably would have chosen Dialogue 2 because it's more matter-of-fact apart from that.
8	Sheldon is verbose, analytical, precise and socially awkward. D1 is considerably more verbose than D2. "Come on" is not a usage I'd expect out of Sheldon. Neither is "oh."
9	Dialogue 1 has more of the complex and quirky tone that Sheldon typically conveys.
10	I think the S2 from Dialogue 1 comes off as more deadpan and confident than dialogue 2.
11	This was a harder decision to make. The shorter sentences of Dialogue 2 are more like Sheldon, but the use of the word "damn" doesn't seem to fit with his speech patterns. On the other hand, dialogue 1 uses more complex sentences than the examples of Sheldon. Overall I chose dialogue 1 because I feel Sheldon is more likely to interject something like "quite unfortunately" than "damn" or "come on."
	Sheldon vs. Raj: Raj-modeled dialogue more similar to Sheldon
12	The "err" phrases in Dialogue 1 do not seem like Sheldon, as he talks very quickly and smoothly without much hesitation and pause.
13	Neither one of these dialogues really stuck out to me as being representative of Sheldon, but I would lean more towards Dialgoue 2. Although the swearing isn't like him, the sentence structure seems to match his patterns more.
14	S2 says "obviously" and "everybody knows that" in Dialogue 2. That seems more in line with Sheldon, who seems like he's not concerned with sparing people's feelings if they don't know as much as he does.
15	I can picture him saying "I thought everybody knew that," it just fits for him perfectly. Its slight condescending without meaning to be.
16	I had a tough time deciding between these two dialogues. The first one seems to suit Sheldon perfectly, but the second line of Dialogue 2 really makes me think of Sheldon. He has an almost feminine or flamboyant way of speaking at times, and I can see that in the "Oh gosh" and "Oh the very oppressive..." In the third line of Dialogue 2, "Come on..." sounds like the arrogant side of Sheldon.
17	I could imagine a Sheldon-like character saying "gosh" as opposed to "God" like in Dialogue 2.
18	When excited, Sheldon can become animated in a quirky way. The lines of Dialogue 2 seem to fit this. I can picture him in his flamboyant way, getting worked up and speaking those lines.
19	Sheldon is verbose, analytical, precise and socially awkward. As previously, I don't like that "as it were" for Sheldon in D1, or the flourish of "unfortunately" in several sentences. D2's S2 seems to go on at length, but in a way that is precise (even though I don't like that "quite oh").
20	I would pick 2 because it is very straightforward and there is no extra side comments.
21	In Dialogue 1, the speaker asks the other person, "can you tell the next part?" Sheldon would not usually ask this, and would just explain it all himself. Dialogue 2 is closer to his character.

Table .8: Penny (in comparison to Leonard) Full Worker Comments on Dialogue

yellow for Penny traits; green for less-Penny (Leonard) traits

	Penny vs. Leonard: Penny-modeled dialogue more similar to Penny
1	There is better wording in Dialogue 1 to give the character more personality. Dialogue 2 is just too monotone for Penny.
2	Dialogue 2 is way to boring to belong to her, I would expect much more randomness for it to fit her character.
3	I think Penny is very talkative. She'd add more commentary than I see from S2 in Dialogue 2, where the language is very direct and bare.
4	I was going to choose Dialogue 2 but I felt there wasn't enough questioning going on. In one where he says magic, huh? Fits much better for how he would ask a question.
5	Penny seems like a character who would seek feedback from other people about things. I see that more from S2 in Dialogue 1 than in Dialogue 2. S2 says things like, "..magic, huh?"
6	Penny comes across as somewhat naive and uneducated, but good hearted. She frequently uses colloquial language, interjections and filler words. D1 uses a good mix of colloquialisms and Penny-like filler. D2 uses some colloquialisms and Penny-like interjections, but there's too much adverb usage going on for it to be a reflection of Penny's speaking style.
7	I almost chose two but 1 still fits better with the random pauses and mhmms.
8	The statements in Dialogue 2 are very straightforward and matter-of-fact. I think Penny would be more like S2 in the first dialogue because she seems like she's not always sure of what she's saying. She'd be likely to say things with hesitation like, "let's see, that..., the crow needed what?"
9	Penny comes across as somewhat naive and uneducated, but good hearted. She frequently uses colloquial language, interjections and filler words. D2 isn't bad for Penny, but "technically" doesn't sound like her. Adverbs that tend to be keyed on precision or intellect usually cause me to rule the dialogue out for her.
10	Saying "everybody knows that" and "obviously" like S2 in Dialogue 2 sounds a little rude, and Penny seems like she really wants people to like her in the sample. I think she'd be nicer than that, like S2 in Dialogue 1.
11	Dialogue 2 sounds a little more natural and simple in this pair, but I still think that Dialogue 1 sounds like an average speaker like Penny.
12	Dialogue 1 has a lot of randomness to it which fits her character really well. The use of, "oh gosh," " I mean," "damn yeah," all of those sound like something she would say.
13	Dialogue 2 is too simple and boring for Penny, for the most part. The speech is mostly monotone, with some more complex words and phrases that I wouldn't normally think of Penny saying.
14	Speaker asks a lot of question, similar to Penny, who often can't keep up with the guys on the show.
15	Dialogue 2 is too matter of fact manner of speaking to be Penny.
16	The lines in Dialogue 2 are too boxy and bland for Penny. She usually has a lot of personality in her speech.
17	I picked dialogue 1 again because in dialogue 2, the 3rd sentence where the person says "where technically" just doesn't sound so much like Penny.
18	Penny comes across as somewhat naive and uneducated, but good hearted. She frequently uses colloquial language, interjections and filler words. Do I see any Penny in D2? There's the "yeah," the higher than usual proportion of questions, and the brief, fairly simple statements. Is it enough to constitute a personality, however? I'm going to say so, although it did require some reflection. D1, as I've said before, at least captures Penny's bubbly cheerfulness."
19	Although I don't think she would use the word darn, this sounds more like her then Dialogue 2 does because 2 is to straight to the point.
20	Penny seems really talkative, and she says "oh" all the time in the sample dialogue, so I think S2 in Dialogue one is more her style of speaking.
	Penny vs. Leonard: Leonard-modeled dialogue more similar to Penny
1	Penny is a very simple speaker, matching Dialogue 2 much better.

Table .9: Penny (in comparison to Bernadette) Full Worker Comments on Dialogue

yellow for Penny traits; green for less-Penny (Bernadette) traits

	Penny vs. Bernadette: Penny-modeled dialogue more similar to Penny
1	I cannot picture Penny starting a sentence with "Essentially" as in Dialogue 2. Dialogue 1 is more natural for Penny as it's more similar to how an average person speaks.
2	I really don't see her saying the words essentially, it seems not her style.
3	I don't think Penny would say "particularly" or "essentially" like S2 does in Dialogue 2. She's more likely to throw an "oh" or "okay" in a sentence like S2 in Dialogue 1.
4	There is to many big words used in dialogue two to really fit with her speech pattern.
5	Dialogue 1 sounds much more like Penny, very simply and short.
6	Penny comes across as somewhat naive and uneducated, but good hearted. She frequently uses colloquial language, interjections and filler words. Both of these have the fingerprints of Penny on them. D2 captures her good cheer in a way that D1 doesn't, but D1 DOES have those repeated usages of "Oh" and "Oh God" which are all over the sample dialogue. If D2 had a few more of those, I could choose it in a heartbeat, but as it stands, D1 has to win by a hair.
7	These dialogues are more similar than the others for Penny for this story, but S2 sounds more silly or uninformed in Dialogue 1, and Penny seems ditzy in the sample. Dialogue 2 has S2 using a more explanatory style, too, which feels unlike Penny.
8	Dialogue 1 has a speaker that isn't completely dumb, but speaks in a natural tone, not too full of complexity. This sounds much more like Penny.
9	I dont really see her saying, "somewhat magic," that really doesn't fit for the way she would talk.
10	Penny talks like a typical young woman, which is much similar to Dialogue 1. The questions, hesitations, and extra words and simple format make this dialogue very similar to what I think Penny would say.
11	I don't think Penny would use the term essentially in a sentence when it was not needed.
12	The conversational and easy lines from Dialogue 1 are a much better fit for Penny than the stuffy and complex lines from Dialogue 2.

	Penny vs. Bernadette: Bernadette-modeled dialogue more similar to Penny
1	Dialogue 2 sounds somewhat sarcastic which fits his speech pattern really well. Dialogue 2 almost sounds perfect but the sarcasm isn't there.
2	In this set, S2 seems more bewildered in Dialogue 2, saying things like, "I suppose," and "Wait..." Penny also seems more likely to use "terrible" to describe something than "erroneous."
3	I could picture Penny getting very animated while saying the lines of Dialogue 2. There is a lot more character in those lines.
4	Penny seems like she thinks while she's talking, and Dialogue 2 shows a little more of that. S2 says things like, "Hold on...who came and stood under the tree?" but she's not asking another person, just trying to figure it out for herself. These two feel similar to me, though.
5	Injection of phrase "you know" in dialogue 2 fits with Penny's informal style of speech.
6	I picked dialogue 2 because most of it definitely sounds as simplistic as Penny talks. Nothing really quirky but the words sound like the words that she would use mostly overall.
7	Penny comes across as somewhat naive and uneducated, but good hearted. She frequently uses colloquial language, interjections and filler words. There's some similarity with these texts, e.g. 1 uses "darn nice" and 2 uses "wonderful." 2 also has Penny's signature interrogative unspeak "right" and "you know." Honestly, these are close, and I tend to be reluctant to favor 2, but I think 2 might reflect her a LITTLE bit more.
8	The strange wording and the use of questions makes me think its more her speech pattern in Dialogue 2.
9	Penny seems to ramble a bit or volunteer information without being asked. Both of the dialogues read like that for me, though. I chose Dialogue 2 because S2 asks more questions, and Penny does that in the sample dialogue, even rhetorical ones.

.7 Conversation Features Analysis

In this section we briefly analyze how characters speak differently to different people in terms of their features in the Z-models. For simplicity we focus on 2-person dialogue only. We collect scenes with dialogue involving only two characters and create corresponding Z-models for each speaker-addressee pair (e.g., Sheldon-Leonard, Amy-Penny). The character that speaks first in the scene is the **speaker**, and the other character is the **addressee**. We want to see if there are features from Z-models that are specific to each pair, and features that stay the same for the speaker regardless of addressee. This provides a deeper insight into the differences in linguistic styles between different speaker and addressee that can be useful for producing target NLG for future work. Note that while we make a distinction about who speaks first in this experiment, for future work we plan to simply combine dialogue with the same two people.

Conversation Pairs for Z=1 Models

Table .10 shows common features for all characters (except for Penny and Leonard) regardless of his/her addressee for Z=1 models. Amy retains the most number of features across different addressees: 8. These include average content word length, six-letter-words, words/utterances per sentence, dialogue act of clarify at the beginning and/or end of the sentence, and the phrase *kind of*. Some of these features have negative Z values depending on the addressee. For example when talking to Leonard, Amy does not use dialogue-act Clarify phrases at the beginning of the utterance, nor does she use the phrase "kind of".

Bernadette retains 4 features across different addressees. These include using the phrases *for* and *kind of* (all negative Z value), conceptual words, and LIWC category Inhibition. Again, some Z values are negative depending on the ad-

dressee. Raj retains 3 features including the phrase *kind of* (all negative Z value) and positive concessive clauses starting with *all* and*but*.

Sheldon and Howard both retain only 1 feature. For Sheldon it is to use LESS dictionary words for all his addressee. For Howard he uses the phrase *kind of* to everyone except Leonard and Sheldon. And finally, Penny and Leonard have retain NO features across addressee.

Table .10: Common Features of Speakers Regardless of Addressee (Z=1 Model)

Speaker	**All Addressee**	**Num Feats**	**Features**
Amy	Penny, Sheldon, Leonard	8	avg-content-wlen, LIWC-SixLetterWords, LIWC-WordsPerSentence, dialact-first-Clarify-ratio (- for Leonard), dialact-last-Clarify-ratio, *kind of* (- for Leonard), tokenspersent, wordspersent
Sheldon	Raj, Penny, Howard, Leonard, Amy	1	LIWC-DictionaryWords(-)
Howard	Bernadette, Raj, Penny, Leonard, Sheldon	1	*kind of* (- for Leonard and Sheldon)
Raj	Bernadette, Penny, Sheldon, Howard, Leonard	3	concess-pol-all-pos (- for Bernadette, Leonard), *kind of* (-), concess-pol-but-pos (- for Bernadette and Leonard)
Bernadette	Raj, Penny, Howard	4	*for* (- for Raj and Penny), *kind of* (-), allwords-conceptverbsperword (- for Raj), LIWC-Inhibition (- for Raj and Penny)

Conversation Pairs for Z=2 Models

Table .11 shows the available Z=2 Models for different pairs. We distinguish the difference between pairs A-B and B-A (e.g., Leonard-Penny, Penny-Leonard) as the person who speaks first often set the tone and content of the conversation.

Table .11: Speaker-Addressee Conversation Pairs for Z=2 Model

	Addressee						
Speaker	Sheldon	Leonard	Penny	Howard	Raj	Bernadette	Amy
Sheldon	self	✓	✓	✓	✓	n/a	✓
Leonard	✓	self	✓	✓	✓	✓	n/a
Penny	✓	✓	self	✓	✓	✓	✓
Howard	✓	✓	✓	self	✓	✓	n/a
Raj	✓	✓	✓	✓	self	✓	n/a
Bernadette	n/a	n/a	✓	✓	✓	self	n/a
Amy	✓	✓	✓	n/a	n/a	n/a	self

Unfortunately there were no common features for characters regardless of his/her addressees for the Z=2 models. However there were some for Z=1 models (contains more features than Z=2 models) which we will discuss later.

Here we look at two speakers A and B, and find common features among A-B and B-A. This indicates **features that stayed the same regardless of who speaks first**. Table .14 shows the two speakers, feature counts, and sample features. This shows some interesting phrase usages that are only specific between two people (see highlighted areas in the table). For example, Sheldon and Leonard talks about *theoretical physics*, Sheldon and Penny has *pal* and *typical*, Penny and Amy has *bitch*, and Howard and Raj has *ninja* in their conversations. Also, the two couples, Sheldon-Amy and Howard-Bernadette, both have the word *wonderful* in their conversation. These additional information gives us a glimpse of characters' intimate relations with each other and may be useful for future work where we want to generate speaker-addressee-specific dialogue.

Another character analysis we performed was to look at **common features with two addressees** (Table .13). Note that this is NOT a multi-party conversation. For example, for Sheldon as a speaker and {Penny, Raj} as addressees, we are comparing Sheldon-Penny and Sheldon-Raj's Z-models. The purpose is to see whether there is something in common when a character talks to two differ-

Table .12: Features of Conversation Pairs Regardless of Who Speaks First

AB = Speaker A to Addressee B
BA = Speaker B to Addressee A
$\cap$ = AB $\cap$ BA
rep-vb-X = repeating verb X within a sentence
dialact = dialogue act

A	**B**	**Num Feats AB : BA : $\cap$**	**Sample Features**
Sheldon	Leonard	106:55:8	*now, already, even if, even though, theoretical physics, though*
	Penny	107:52:12	*rather, pal, right, where, you know, so, following, typical, rather*
	Howard	33:39:2	"*you,*", verb-strength
	Raj	30:32:7	*however, third*, emotion words
	Amy	49:35:7	*while, wonderful*, repeat-verb-*think*, *terrible*, *ass*
Penny	Leonard	51:70:12	*worst, even though, who, even if, just, kind of*, rep-verb-*wait*
	Raj*	55:35:3	*better*, dialact-emphasis, dialact-accept
	Howard	34:37:5	*oh, kind of*, dialact-greet
	Bernadette	40:82:4	*what*, concept verbs
	Amy	33:41:2	*bitch*
Leonard	Howard	46:40:3	*oh, believe*
	Raj	39:37:2	LIWC-future-tense
	Amy	45:39:2	*who*
Howard	Raj	36:35:2	"! You", "ninja,"
	Bernadette	35:35:4	rep-vb-*get*, *wonderful*, *to live*
Bernadette	Raj*	56:66:7	LIWC-certainty, *nice, that*

* For readers that know BBT well, for a while Raj could not talk to girls unless he (thinks he) is drunk. However he fantasizes conversation with the girls in some scenes.

ent addressees separately. There were some interesting character analysis, such as Sheldon using *knock-knock-knock* to both Penny and Amy, and Howard using emotional phrases with Sheldon and Leonard. However this does not seem to contribute much for NLG as Table .14 provides sufficient information already.

Features specific to a speaker-addressee pair (who speaks first matters) is shown in Tables .14, .15, and .16. The overlapping features are removed from each speaker's different addressee. For example, Sheldon==Leonard (speaker is Sheldon; addressee is Leonard) has 106 features in its Z=2 model. We remove

Table .13: Common Features with Two Addressee Separately

concess-pol-W-P = concessive clause word W and polarity P (obj = neutral)
(-) = negative Z score

Speaker	Addressee	Num Feats	Sample Features
Sheldon	Penny, Raj	1	“, which"
	Penny, Howard	1	*also*
	Penny, Leonard	41	*rather, nor, somewhat*, concess-pol-while-obj, concess-pol-however-neg
	Penny, Amy	7	rep-vb-knock, *actually*, concess-pol-*though*-obj
	Leonard, Howard	1	*whom*
	Leonard, Amy	5	concess-pol-*while*-neg, concess-pol-*though*-obj, rep-vb-sit,
	Amy, Howard	1	interjection followed by a period
Penny	Sheldon, Leonard	9	rep-vb-wait, rep-vb-read, *worst, just, so*
	Raj, Amy	1	*what happened*
	Raj, Howard	5	dialact-first-greet, *with, kind of*, LIWC-ThirdPersonSingular
	Raj, Bernadette	3	LIWC-articles, number of hedges per utterance, *kind of*
	Howard, Bernadette	4	concept-adverb-per-word, LIWC-present-tense, *kind of*
Leonard	Sheldon, Howard	1	concess-pol-*while*-pos
	Sheldon, Penny	13	rep-vb-ask, *bad, worse, good, even though*, concess-pol-while-pos, concess-pol-even-if-pos
	Raj, Howard	1	dialact-last-statement(-)
	Raj, Amy	2	“’m not", *as*
	Howard, Penny	1	concess-pol-*while*-pos
	Howard, Amy	1	dialact-last-continuer
Howard	Sheldon, Leonard	3	*by, later*, dialact-emotion
	Raj, Bernadette	4	concess-pol-yet-obj, *least, particularly*
	Raj, Penny	2	*then*
	Penny, Leonard	3	dialact-greet, *kind of*
Raj	Sheldon, Howard	4	rep-vb-want, rep-vb-go
	Penny, Bernadette	2	dialact-first-reject; LIWC-FirstPersonSingular
	Penny, Leonard	3	noun phrase followed by period, TO-verb phrase, *best*
	Leonard, Bernadette	1	LIWC-discrepancy
Amy	Sheldon, Penny	3	rep-vb-go
	Sheldon, Leonard	1	*wonderful*

overlapping features with Sheldon==Penny and left with 65 features. Repeat the process for remaining addressee and we are left with 61 features.

Similar to Table .14, the results give us insights on how characters talk differently to different people in terms of their. From previous experiments we know

that Sheldon says phrases such as *literally, quintessential, typical*. Here we have a better idea on **to whom** the phrases are used: *literally* (to Amy), *quintessential* (to Leonard), and *typical* (to Penny). In addition, tag-question is not a significant feature for Sheldon overall, but it is significant when talking to Howard. Leonard as the speaker also shows some interesting outcome. For example, he shows some conflicting emotions to Amy, as we see him use both negative and positive emotion words. A dialogue act of Accept is used with Penny, perhaps an indication of his friendliness towards her. When talking to Sheldon he likes to repeat the same verb in the same sentence: *go, think, lie, operate, take, make, read*.

Table .14: Features Specific to Speaker-Addressee Pairs

Speaker	Addressee	Num Feats	Sample Features
Sheldon	Amy	38	*buddy, ass, to stop, terrible, following, at the same time, often, literally*, rep-vb-think, dialact-last-Bye,
	Howard	30	*now, first, until, also, must, until*, "you,", rep-vb-please, verb-strength, tag-ratio
	Leonard	61	*theoretical physics, bitch, sucks, quintessential*, rep-vb-{need, buy, complain, accept}, concess-pol-even-if-pos, concess-pol-despite-pos,
	Penny	59	*typical, which, actually, sort of, whereas, it seems*, concess-pol-yet-neg, dialact-bye, rep-vb-control
	Raj	29	*yet, however, less, earlier, third, very*, concess-pol-yet-obj, dialact-first-Continuer
Leonard	Amy	42	*who, seem, just, really, you're, as, so*, neg polarity overall, LIWC-PositiveEmotion, emotion words, LIWC-discrepancy(-), LIWC-exclusive, dialact-continuer
	Howard	43	*next, with, while, what happened, oh, believe, you have, good, while, around, next, why, very*, ", there", ", with", overall rep-vb, merge-ratio, rep-vb-{do, come, have}, dialact-first-whQuestion
	Penny	57	*sucks, wonderful, buddy, basically, terrible, least, around, really, when, right, who, during, just, nice, first, worse, kind of, I get, but, actually, yeah*, "where's", concess-pol-though-neg, concess-pol-yet-neg, rep-vb-{regret, wait, love, look, view}, dialact-accept
	Raj	36	*yeah, beautiful, and, with*, rep-vb-stay, LIWC-{QuestionMarks, FutureTense, Articles(-), Inhibition, Inclusive, Conjunctions, Assent}, dialact-first-continuer, dialact-ynQuestion, dialact-yAnswer, dialact-statement-ratio(-)
	Sheldon	42	*nevertheless, pretty, should, already, quite, hell, feel, even though, now, why, theoretical physics*, rep-vb-{go, think, lie, operate, take, make, read}, concess-pol-though-obj, concess-pol-even-though-obj

Table .15: Features Specific to Speaker-Addressee Pairs (cont.)

Speaker	Addressee	Num Feats	Sample Features
Penny	Amy	32	*should, pretty, second, bitch, quite*, rep-vb-{tell, get}, LIWC-assent
	Bernadette	76	*but, gonna, right, when, kind of, because*, avg-content-wlen(-), concession ratio, LIWC-{Feel, Impersonal-Pronouns, DictionaryWords, Quantifiers, Prepositions, CommonVerbs, SixLetterWords(-), Conjunctions, Adverbs, Causation, Tentative, TotalFunctionWords, Anger}, dialact-{reject(-), ynQuestion, statement}, concess-pol-all-obj, concess-pol-but-obj
	Howard	36	*why, first, by, you know, with*, "oh,", polarity-percent-neg(-), polarity-percent-pos, LIWC-{Discrepancy(-), Non-fluencies, PersonalPronouns}, rep-vb-APOSTROPHEm, dialact-{Greet, yAnswer}
	Leonard	42	*even though, later, damn, since, then, worst, kind of, hell, ass, next, even if*, rep-vb-{think, take, cry}, concess-pol-yet-obj, concess-pol-even-if-pos, concess-pol-though-pos
	Raj	47	*last, just, better, most, you have, since, better*, words per utterance, LIWC-{ThirdPersonPlural, Insight}, verb repetition ratio, merge-ratio, verb-strength(-), rep-vb-{know, come, say, mean, APOSTROPHEs}, dialact-{clarify, first-accept, first-emphasis}
	Sheldon	43	*typical, pal, rather*, concess-pol-on-the-other-hand, rep-vb-{need, care, try, allow, die}
Amy	Leonard	38	*who, beautiful, but, hear you, because, most, the same, next, very*, LIWC-{ThirdPersonPlural, SixLetterWords, Anxiety} rep-vb-{be, APOSTROPHEm}, avg-content-wlen, dialact-{first-nAnswer, first-Reject}
	Penny	38	*during, bitch, best, following, on the other hand, particularly, often, literally*, in-group words, dialact-first-continuer, rep-vb-go
	Sheldon	31	*while, terrible, ass, worst, technically, almost*, ", which", rep-vb-{kiss, think}, dialact-first-clarify,
Bernadette	Howard	35	*wonderful, though, finally, during, or, how to, I get, third, almost*, rep-vb-{make, get}, "honeymoon."
	Penny	40	*too bad, and, where, already, as*, LIWC-{Sadness, Anxiety, Fillers, See}, dialact-{Continuer, last-Reject}, RID-primary, RIDsecondary(-), polarity-overall (-)
	Raj	56	*nice, damn right, with, hell, that*, LIWC-{Certainty, FirstPersonPlural, AuxiliaryVerbs, SecondPerson, NegativeEmotion, Anger, TotalFunctionWords, TotalPronouns, Tentative}, dialact-last-Clarify, persuasive words, near-swear words, dialact-{whQuestion, emphasis}

Table .16: Features Specific to Speaker-Addressee Pairs (cont.)

Speaker	Addressee	Num Feats	Sample Features
Howard	Bernadette	31	*wonderful, roughly, heck, worse, earlier, beautiful, while, soon*, rep-vb-{float, get}
	Leonard	34	*more, who, last, oh, later, believe, or, better*, persuasive words, LIWC-Prepositions, rep-vb-believe, dialact-{emotion, reject, greet}
	Penny	32	*beautiful, or, to see*, "oh,", conjunctions, LIWC-{hear, feel, exclusive}, rep-vb-{do, speak}, dialact-nAnswer
	Raj	30	*pal, whose, second, yet, which*, in-group words, dialact-bye
	Sheldon	36	*for, yeah, ass, during, nice, by, at the same time, around, so, really*, verb-strength, polarity-percent-obj, LIWC-See, concess-pol-but-neg, rep-vb-{say, APOSTROPHEs}
Raj	Bernadette	63	*so, think, with*, LIWC-{Certainty, PositiveEmotion, Nonfluencies, CommonVerbs, Insight, Negations(-), PersonalPronouns, Questionmarks, Quantifiers(-), PastTense, Exclusive(-)}, dialact-{ynQuestion, statement, reject(-)}, category-agg, opinion words, justify words, emotion words
	Howard	31	*finally, her?, heck, actually, soon*, "ninja,", concess-pol-despite-pos
	Leonard	33	*more, first, where, best, really, first*, "Oh,", polarity-sents(-), LIWC-{futuretense, fillers}, dialact-{last-accept, emphasis}, concess-pol-but-neg, concess-pol-all-neg, tag questions, rep-vb-have,
	Penny	30	*now, better, good, should, very*, "Sorry.", concess-pol-but-pos, LIWC-sadness, RIDseconday(-), dialact-{emphasis, first-accept},
	Sheldon	28	*however, until, or, sort of, third*, concess-pol-however-pos, rep-vb-go

Bibliography

[Allport, 1960] Allport, G. (1960). *Personality and social encounter: Selected essays.* Beacon.

[Allport and Odbert, 1936] Allport, G. and Odbert, H. (1936). Trait-names: A psycho-lexical study. *Psychological monographs*, 47(1):i.

[Ameixa et al., 2013] Ameixa, D., Coheur, L., and Redol, R. A. (2013). From subtitles to human interactions: introducing the subtle corpus. Technical report, IST/INESC-ID.

[Baccianella et al., 2010] Baccianella, S., Esuli, A., and Sebastiani, F. (2010). Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), Valletta, Malta, May. European Language Resources Association (ELRA).* Avaliable at http://sentiwordnet.isti.cnr.it/.

[Bamman et al., 2014a] Bamman, D., O'Connor, B., and Smith, N. A. (2014a). Learning latent personas of film characters. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, page 352.

[Bamman et al., 2014b] Bamman, D., Underwood, T., and Smith, N. A. (2014b). A bayesian mixed effects model of literary character. In *ACL (1)*, pages 370–379.

[Banchs, 2012] Banchs, R. E. (2012). Movie-dic: a movie dialogue corpus for research and development. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 203–207. Association for Computational Linguistics.

[Bednarek, 2011] Bednarek, M. (2011). The language of fictional television: a case study of the dramedy gilmore girls. *English Text Construction*, 4(1):54–83.

[Bednarek, 2012] Bednarek, M. (2012). Constructing "nerdiness": Characterisation in the big bang theory. *Multilingua*, 31(2-3):199–229.

[Bee et al., 2010] Bee, N., Pollock, C., André, E., and Walker, M. (2010). Bossy or wimpy: expressing social dominance by combining gaze and linguistic behaviors. In *Intelligent Virtual Agents*, pages 265–271. Springer.

[Berne, 1996] Berne, E. (1996). *Games people play: The psychology of human relationships*. Tantor eBooks.

[Bollacker et al., 2008] Bollacker, K., Evans, C., Paritosh, P., Sturge, T., and Taylor, J. (2008). Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250. ACM.

[Bouayad-Agha et al., 1998] Bouayad-Agha, N., Scott, D., and Power, R. (1998). Integrating content and style in documents: a case study of patient information leaflets. *Information design journal*, 9(2-3):2–3.

[Bowden, 2016] Bowden, K. (2016). PyPer. In progress.

[Bredin et al., 2014] Bredin, H., Roy, A., Pécheux, N., and Allauzen, A. (2014). Sheldon speaking, bonjour!: Leveraging multilingual tracks for (weakly) supervised speaker identification. In *Proceedings of the ACM International Conference on Multimedia*, pages 137–146. ACM.

[Brown and Levinson, 1987] Brown, P. and Levinson, S. C. (1987). *Politeness: Some universals in language usage*, volume 4. Cambridge university press.

[Bubel, 2005] Bubel, C. (2005). *The linguistic construction of character relations in TV drama: Doing friendship in Sex and the City*. PhD thesis, Universität des Saarlandes.

[Callaway and Lester, 2002] Callaway, C. B. and Lester, J. C. (2002). Narrative prose generation. *Artificial Intelligence*, 139(2):213–252.

[Cassell et al., 1999] Cassell, J., Torres, O., and Prevost, S. (1999). Turn taking vs. discourse structure: How best to model multimodal conversation. *Machine conversations*, pages 143–154.

[Cavazza and Charles, 2005] Cavazza, M. and Charles, F. (2005). Dialogue generation in character-based interactive storytelling. In *AIIDE*, pages 21–26.

[Danescu-Niculescu-Mizil and Lee, 2011] Danescu-Niculescu-Mizil, C. and Lee, L. (2011). Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, ACL 2011*.

[Davies, 2012a] Davies, M. (2012a). Comparing the corpus of american soap operas, coca, and the bnc. `http://corpus.byu.edu/soap/overview.asp`.

[Davies, 2012b] Davies, M. (2012b). Corpus of american soap operas. `http://corpus.byu.edu/soap/`.

[Dunbar, 1998] Dunbar, R. (1998). *Grooming, gossip, and the evolution of language*. Harvard University Press.

[Elson and McKeown, 2009] Elson, D. and McKeown, K. (2009). A tool for deep semantic encoding of narrative texts. In *Proceedings of the ACL-IJCNLP 2009 Software Demonstrations*, pages 9–12. Association for Computational Linguistics.

[Fiske, 2002] Fiske, J. (2002). *Television culture*. Routledge.

[Forsyth and Martell, 2007] Forsyth, E. and Martell, C. (2007). Lexical and discourse analysis of online chat dialog. In *Semantic Computing, 2007. ICSC 2007. International Conference on*, pages 19–26. IEEE.

[Furnham, 1990] Furnham, A. (1990). Language and personality. *Handbook of Language and Social Psychology*.

[Goffman, 1970] Goffman, E. (1970). *Strategic interaction*, volume 1. University of Pennsylvania Press.

[Goldberg, 1990] Goldberg, L. (1990). An alternative" description of personality": the big-five factor structure. *Journal of Personality and Social Psychology; Journal of Personality and Social Psychology*, 59(6):1216.

[Gordon et al., 2007] Gordon, A. S., Cao, Q., and Swanson, R. (2007). Automated story capture from internet weblogs. In *Proceedings of the 4th international conference on Knowledge capture*, pages 167–168. ACM.

[Gupta et al., 2007] Gupta, S., Walker, M. A., and Romano, D. M. (2007). How rude are you?: Evaluating politeness and affect in interaction. In *Affective Computing and Intelligent Interaction*, pages 203–217. Springer.

[Guzdial et al., 2014] Guzdial, M., Harrison, B., Li, B., and Riedl, M. O. (2014). Crowdsourcing open interactive narrative. In *the 10th International Conference on the Foundations of Digital Games. Pacific Grove, CA*.

[Hall et al., 2009] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

[Ireland et al., 2011] Ireland, M. E., Slatcher, R. B., Eastwick, P. W., Scissors, L. E., Finkel, E. J., and Pennebaker, J. W. (2011). Language style matching predicts relationship initiation and stability. *Psychological Science*, 22(1):39–44.

[Kacewicz et al., 2011] Kacewicz, E., Pennebaker, J., Davis, M., Jeon, M., and Graesser, A. (2011). Pronoun use reflects standings in social hierarchies. *Submitted for publication*.

[Kozloff, 2000] Kozloff, S. (2000). *Overhearing film dialogue*. University of California Press.

[Labov, 2006] Labov, W. (2006). *The social stratification of English in New York city*. Cambridge University Press.

[Li et al., 2012] Li, B., Lee-Urban, S., Appling, D. S., and Riedl, M. O. (2012). Crowdsourcing narrative intelligence. *Advances in Cognitive Systems*, 2:25–42.

[Li et al., 2013] Li, B., Lee-Urban, S., Johnston, G., and Riedl, M. (2013). Story generation with crowdsourced plot graphs. In *AAAI*.

[Li et al., 2014] Li, B., Thakkar, M., Wang, Y., and Riedl, M. O. (2014). Data-driven alibi story telling for social believability. In *Proceedings of the FDG 2014 Social Believability in Games Workshop*. Citeseer.

[Lin and Walker, 2011a] Lin, G. and Walker, M. (2011a). All the world's a stage: Learning character models from film. *Proceedings of the Seventh AI and Interactive Digital Entertainment Conference*.

[Lin and Walker, 2011b] Lin, G. and Walker, M. (2011b). All the world's a stage: Learning character models from film. In *Proceedings of the Seventh AI and Interactive Digital Entertainment Conference, AIIDE*, volume 11.

[Lukin et al., 2014] Lukin, S. M., Ryan, J. O., and Walker, M. A. (2014). Automating direct speech variations in stories and games. In *Tenth Artificial Intelligence and Interactive Digital Entertainment Conference*.

[Lukin and Walker, 2015] Lukin, S. M. and Walker, M. A. (2015). Narrative variations in a virtual storyteller. In *Intelligent Virtual Agents*, pages 320–331. Springer.

[Mairesse and Walker, 2011] Mairesse, F. and Walker, M. (2011). Controlling user perceptions of linguistic style: Trainable generation of personality traits. *Computational Linguistics*.

[Mairesse et al., 2007] Mairesse, F., Walker, M., Mehl, M., and Moore, R. (2007). Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research (JAIR)*, 30:457–500.

[Mairesse and Walker, 2010] Mairesse, F. and Walker, M. A. (2010). Towards personality-based user adaptation: psychologically informed stylistic language generation. *User Modeling and User-Adapted Interaction*, 20(3):227–278.

[Mateas, 2007] Mateas, M. (2007). The authoring bottleneck in creating ai-based interactive stories. In *Proceedings of the AAAI 2007 Fall Symposium on Intelligent Narrative Technologies*.

[Mateas and Stern, 2003] Mateas, M. and Stern, A. (2003). Façade: An experiment in building a fully-realized interactive drama. In *Game Developers Conference, Game Design track*, volume 2, page 82.

[McCoy et al., 2012] McCoy, J., Treanor, M., Samuel, B., Reed, A. A., Wardrip-Fruin, N., and Mateas, M. (2012). Prom week. In *Proceedings of the International Conference on the Foundations of Digital Games*, pages 235–237. ACM.

[McCoy et al., 2011] McCoy, J., Treanor, M., Samuel, B., Wardrip, N., and Mateas, M. (2011). Comme il faut: A system for authoring playable social models. In *Seventh Artificial Intelligence and Interactive Digital Entertainment Conference*.

[Meehan, 1977] Meehan, J. R. (1977). Tale-spin, an interactive program that writes stories. In *Proceedings of the Fifth International Conference on Artificial Intelligence (IJCAI 77)*.

[Mott and Lester, 2006] Mott, B. and Lester, J. (2006). Narrative-centered tutorial planning for inquiry-based learning environments. In *Intelligent Tutoring Systems*, pages 675–684. Springer.

[Munteanu et al., 2010] Munteanu, A., Costea, I., Palos, R., and Jinaru, A. (2010). Entering in the essences of personality–studies over archetypes. *Procedia-Social and Behavioral Sciences*, 5:2272–2276.

[Neff et al., 2011] Neff, M., Toothman, N., Bowmani, R., Fox Tree, J., and Walker, M. (2011). Don't scratch! self-adaptors reflect emotional stability. In *Intelligent Virtual Agents*, pages 398–411. Springer.

[Neff et al., 2010] Neff, M., Wang, Y., Abbott, R., and Walker, M. (2010). Evaluating the effect of gesture and language on personality perception in conversational agents. In *Intelligent Virtual Agents*, pages 222–235. Springer.

[Neviarouskaya et al., 2009] Neviarouskaya, A., Prendinger, H., and Ishizuka, M. (2009). Emoheart: Conveying emotions in second life based on affect sensing from text. *Advances in Human-Computer Interaction*.

[Norman, 1963] Norman, W. (1963). Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *The Journal of Abnormal and Social Psychology*, 66(6):574.

[Oberlander et al., 2000] Oberlander, J., Brew, C., Oberlander, J., and Brew, C. (2000). Stochastic text generation. *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 358(1769):1373–1387.

[Paiva and Evans, 2004] Paiva, D. and Evans, R. (2004). A framework for stylistically controlled generation. *Natural Language Generation*, pages 120–129.

[Paiva and Evans, 2005] Paiva, D. and Evans, R. (2005). Empirically-based control of natural language generation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 58–65. Association for Computational Linguistics.

[Pennebaker and L.A., 1999] Pennebaker, J. and L.A., K. (1999). Linguistic styles: language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296.

[Pennebaker, 2011] Pennebaker, J. W. (2011). *The Secret Life of Pronouns: What Our Words Say About Us*. Bloomsbury Press.

[Pennebaker and Chung, 2012] Pennebaker, J. W. and Chung, C. K. (2012). Language and social dynamics. Technical report, DTIC Document.

[Pennebaker et al., 2001] Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71:2001.

[Pennebaker and Ireland, 2011] Pennebaker, J. W. and Ireland, M. E. (2011). Using literature to understand authors: The case for computerized text analysis. *Scientific Study of Literature*.

[Piwek, 2003] Piwek, P. (2003). A flexible pragmatics-driven language generator for animated agents. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 2*, pages 151–154. Association for Computational Linguistics.

[Quaglio, 2009] Quaglio, P. (2009). *Television Dialogue: The sitcom Friends vs. natural conversation.* John Benjamins Publishing Company.

[Quinlan, 2014] Quinlan, J. R. (2014). *C4. 5: programs for machine learning*. Elsevier.

[Reed et al., 2011a] Reed, A., Samuel, B., Sullivan, A., Grant, R., Grow, A., Lazaro, J., Mahal, J., Kurniawan, S., Walker, M., and Wardrip-Fruin, N. (2011a). Spyfeet: An exercise rpg. *Proceedings of the Sixth International Conference on the Foundations of Digital Games.*, pages 310–312.

[Reed et al., 2011b] Reed, A., Samuel, B., Sullivan, A., Grant, R., Grow, A., Lazaro, J., Mahal, J., Kurniawan, S., Walker, M., and Wardrip-Fruin, N. (2011b). A step towards the future of role-playing games: The spyfeet mobile rpg project. *Proceedings of the Seventh AI and Interactive Digital Entertainment Conference (AIIDE)*.

[Riedl and Young, 2004] Riedl, M. and Young, R. M. (2004). An intent-driven planner for multi-agent story generation. *Proc. of the 3rd Int. Conf. on Autonomous Agents and Multi Agent Systems*.

[Rishes et al., 2013] Rishes, E., Lukin, S. M., Elson, D. K., and Walker, M. A. (2013). Generating different story tellings from semantic representations of narrative. In *Interactive Storytelling*, pages 192–204. Springer.

[Rowe et al., 2008] Rowe, J. P., Ha, E. Y., and Lester, J. C. (2008). Archetype-driven character dialogue generation for interactive narrative. In *Intelligent Virtual Agents*, pages 45–58. Springer.

[Roy et al., 2014] Roy, A., Guinaudeau, C., Bredin, H., and Barras, C. (2014). Tvd: A reproducible and multiply aligned tv series dataset. In *LREC*, pages 418–425.

[Ryan et al., 2014] Ryan, J. O., Barackman, C., Kontje, N., Owen-Milner, T., Walker, M. A., Mateas, M., and Wardrip-Fruin, N. (2014). Combinatorial dialogue authoring. In *Interactive Storytelling*, pages 13–24. Springer.

[Ryan et al., 2015] Ryan, J. O., Fisher, A. M., Owen-Milner, T., Mateas, M., and Wardrip-Fruin, N. (2015). Toward natural language generation by humans. In *Proceedings of the INT*.

[Schmidt, 2007] Schmidt, V. (2007). *45 Master Characters*. Writers Digest Books.

[Schreibman et al., 2008] Schreibman, S., Siemens, R., and Unsworth, J. (2008). *A companion to digital humanities*. John Wiley & Sons.

[Serban et al., 2015a] Serban, I. V., Lowe, R., Charlin, L., and Pineau, J. (2015a). A survey of available corpora for building data-driven dialogue systems. *arXiv preprint arXiv:1512.05742*.

[Serban et al., 2015b] Serban, I. V., Sordoni, A., Bengio, Y., Courville, A., and Pineau, J. (2015b). Hierarchical neural network generative models for movie dialogues. *arXiv preprint arXiv:1507.04808*.

[Shaffer et al., 2005] Shaffer, D., Squire, K., Halverson, R., and Gee, J. (2005). Video games and the future of learning. *WCER Working Paper No. 2005-4*.

[Sienkiewicz et al., 2012] Sienkiewicz, J., Skowron, M., Paltoglou, G., and Holyst, J. (2012). Entropy-growth-based model of emotionally charged online dialogues. *arXiv preprint arXiv:1201.5477*.

[Skorupski et al., 2007] Skorupski, J., Jayapalan, L., Marquez, S., and Mateas, M. (2007). Wide ruled: a friendly interface to author-goal based story generation. *Virtual Storytelling. Using Virtual Reality Technologies for Storytelling*, pages 26–37.

[Stamatatos, 2009] Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556.

[Stolcke et al., 2011] Stolcke, A., Zheng, J., Wang, W., and Abrash, V. (2011). Srilm at sixteen: Update and outlook. In *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop*, page 5.

[Sullivan et al., 2010a] Sullivan, A., Mateas, M., and Wardrip-Fruin, N. (2010a). Rules of engagement: moving beyond combat-based quests. In *Proceedings of the Intelligent Narrative Technologies III Workshop*, page 11. ACM.

[Sullivan et al., 2010b] Sullivan, A., Mateas, M., and Wardrip-Fruin, N. (2010b). Rules of engagement: moving beyond combat-based quests. *Proceedings of the Intelligent Narrative Technologies III Workshop*, pages 1–8.

[Tausczik and Pennebaker, 2010] Tausczik, Y. R. and Pennebaker, J. W. (2010). The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

[Tearse et al., 2010] Tearse, B., Mateas, M., and Wardrip-Fruin, N. (2010). Minstrel remixed: a rational reconstruction. In *Proceedings of the Intelligent Narrative Technologies III Workshop*, page 12. ACM.

[Traum et al., 2007] Traum, D., Roque, A., Leuski, A., Georgiou, P., Gerten, J., Martinovski, B., Narayanan, S., Robinson, S., and Vaswani, A. (2007). Hassan: A virtual human for tactical questioning. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, pages 71–74.

[Turner, 1993] Turner, S. R. (1993). *Minstrel: a computer model of creativity and storytelling*. PhD thesis, University of California at Los Angeles.

[Vinyals and Le, 2015] Vinyals, O. and Le, Q. (2015). A neural conversational model. *arXiv preprint arXiv:1506.05869*.

[Vogel and Lynch, 2008] Vogel, C. and Lynch, G. (2008). Computational stylometry: Who's in a play? *Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction*, pages 169–186.

[Walker et al., 2011a] Walker, M., Grant, R., Sawyer, J., Lin, G., Wardrip-Fruin, N., and Buell, M. (2011a). Perceived or not perceived: Film character models for expressive nlg. *International Conference on Interactive Digital Storytelling*.

[Walker et al., 1997] Walker, M. A., Cahn, J. E., and Whittaker, S. J. (1997). Improvising linguistic style: Social and affective bases for agent personality. In *Proceedings of the first international conference on Autonomous agents*, pages 96–105. ACM.

[Walker et al., 2012] Walker, M. A., Lin, G. I., and Sawyer, J. (2012). An annotated corpus of film dialogue for learning and characterizing character style. In *LREC*, pages 1373–1378.

[Walker et al., 2011b] Walker, M. A., Lin, G. I., Sawyer, J., Grant, R., Buell, M., and Wardrip-Fruin, N. (2011b). Murder in the arboretum: Comparing character models to personality models. In *Workshops at the Seventh Artificial Intelligence and Interactive Digital Entertainment Conference*.

[Walker and Rambow, 2002] Walker, M. A. and Rambow, O. C. (2002). Spoken language generation. *Computer Speech & Language*, 16(3):273–281.

[Wei and Calvert, 2011] Wei, H. and Calvert, T. (2011). Conventions and innovations: Narrative structure and technique in heavy rain. *The Journal of the International Digital Media and Arts Association*, 8:59–68.