Linguistic History Matching Algorithm

Ben Marks

July 25, 2014

1 Introduction

The following document describes the design and implementation process of my Introduction to Linguistics final project: a program to interface with stories in the collected class corpus. I have included elements from all parts of the process, including the initial proposal and design, details about the actual implementation, and a reflection on information from the corpus gained from using the program.

Sections 2 - 6 were written in the early stages of this project. Section 2 describes my initial goals for the project. Section 3 summarizes the key decisions with regard to my algorithm for classifying stories. In section 4, I describe the broad types of questions used to represent a single bilingual story. For this project, I created a series of questions and statements to classify the bilingual experience and then estimated likely responses to each item based on the recounted story in the corpus. Section 5 describes how these estimated answers are used to determine the Cosine similarity between any pair of stories. Section 6 discusses typical approaches to handling gaps in data - a relavent question when classifying the corpus, since some questions may not be answered explicitly in a recount.

Sections 7 - 10 describe the implemented algorithm and program. The algorithm is described in Section 7. Program control flow is described in Section 8, and instructions for running the program are listed in Section 9. Section 10 includes the final list of questions and statements used to classify stories.

Sections 11 - 14 include reflections and insights derived from the implemented program. Section 11 includes a discussion of changes to the algorithm and questions used over the course of the project. Some potential future extension for the existing project are considered in Section 12. The process of creating the working program and classifying the stories led to some interesting observations about the corpus and interview process as described in Section 13. Finally, Section 14 describes some of the aspects of the project that indicate its success.

2 Purpose

The following document outlines the algorithm I will use to estimate the commonality of any pair of stories in the Linguistic Autobiography Corpus constructed by the Introduction to Linguistics students at Swarthmore this spring. The typical use case involves comparing a new story to the stories in the existing database to find stories most similar to the new story.

3 Executive Summary

By quantifying stories on a set of representative characteristics, I hope to facilitate similarity matching between bilingual accounts. A Likert type scale, [1] coupled with quantitative responses is used to quantify stories in a relatively universal way [4]. Each response is paired with a confidence level, indicating to what extent the response is explicitly supported by the linguistic account. Estimations are based on the assumption that the bilingual speaker has been complete in their responses; it is assumed that no data are repressed. The user analyzing the data can choose a minimum threshold of confidence to use when comparing stories. When comparing stories A and B, only questions answered in both at or above the user-selected confidence level will be used. Cosine similarity is used to evaluate the similarity between the set of shared responses. Responses with a similarity closest to one are most similar; responses with a similarity closest to negative one are most dissimilar. The similarity rating is converted to a percentage (0 least similar; 100 most similar) and reported to the user.

4 Data Representation

Each story must be represented in a quantitative way in order to facilitate reasonably accurate comparisons. After skimming through the existing corpus and guiding list of questions given to interviewers, I have developed a series of items - in the form of both questions and statements - that touch on different characteristics of the bilingual experience. The questions address quantitative aspects of the bilingual experience, such as the number of language spoken or the number of languages used each week. The statements address the qualitative aspects of the bilingual experience. Phrased as a remark made by a bilingual speaker, the

[1] Technically, a Likert Scale requires the items to be formatted horizontally, as described in [4]. Therefore, a Likert type scale is a more appropriate description. For the sake of brevity, this technicality will be ignored.

interviewee indicates, on a Likert type scale, to what extent they agree or disagree with the statement as it pertains to their own experience.

Ideally, the questions and statements would be evaluated directly by the bilingual speaker interviewed. Unfortunately, it is infeasible for me to give these questions directly to the interviewees. Therefore, in this project, I will read each story in the corpus and attempt to estimate, for each item, what the bilingual speaker is likely to respond. Handling the subjectivity of these evaluations is discussed in great detail in Sections 6.1 and 6.2.

Together, the quantitative and qualitative items incorporate aspects of language acquisition, experience, use, identity, and intentions, and provide a relatively holistic summary of a speaker's story. The validity of using these particular aspects is supported by the fact that they closely mirror the set of questions given to interviewers, which was collectively constructed and endorsed by the entire LING 01 class. A more thorough discussion of the validity of the questions chosen can be found in Section 13.1.

A list of items, both qualitative and quantitative, is included in Section 10. A detailed discussion of the different considerations made when developing these questions and statements can be found in Section 11.1.

4.1 Qualitative Statements

In reading the corpus, I noticed that certain topics, such as raising children as bilinguals, the context in which languages were learned, or the professional uses of bilingualism, were addressed in some way by the majority of accounts. In order to capture these experiences in a quantitative way, I created a list of statements that allude to the bilingual experience for that person with respect to each noted topic. The statements are intended to be general enough that they apply to most bilinguals, and specific enough to reference a reasonably definitive characteristic, thus allowing for comparison between responses. Further, I attempted to ensure that each category of guiding questions provided to the interviewers had at least one statement addressing that aspect of the bilingual experience.

Some examples of statements include:

- "The language I spoke in primary school is now my dominant language."
- "I wanted to be bilingual and learn the languages I've learned."
- "I choose to address people in the language that I think will make them most comfortable"

Each of these statements is associated with one of six possible responses based on a Likert Scale:

1. Strongly Agree
2. Agree
3. Neutral
4. Disagree
5. Strongly Disagree
6. Question Unanswered

Using a Likert scale offers many benefits in this particular analysis. First, compared to a simple Yes / No response, a Likert scale better captures the diversity of experience by providing a continuum of relative conviction for each response. This accounts for and recognizes differences in intensity of response between stories for any given statement. A Likert scale is also easily quantifiable, meaning different responses can be compared in a relatively straightforward way. Finally, the concepts of agreement and disagreement used by a Likert scale are, for the most part, independent of language, meaning that non-English speakers could understand and respond to the questions in their own language. [2]

In addition to the strength of agreement associated with each statement, each response is paired with one of three levels of confidence. Responses can either have "high confidence," "medium confidence," or "low confidence." Associating different levels of confidence with responses greatly facilitates handling cases where some questions are unanswered. Further, it allows the analyst to have more control over the type of matches returned. Both of these statements are discussed in detail in Section 11.2

[2]There is a fascinating discussion about whether Likert scales suffer from cultural bias. Differences between cultures, especially regarding collectivist vs. individualistic emphases, can influence how survey participants respond to questions on a Likert Scale. It seems these differences are most salient in questions regarding interaction with others. In the context of this algorithm, only select questions involving the speaker's experiences reference interaction with others. More specifically, the questions involving reactions of others to bilingualism and negative effects of bilingualism may suffer from culture-dependent interpretation. The vast majority of questions should be largely independent of cultural interpretation, and comparisons involving a Likert Scale are appropriate. Even with the two questions noted above that may be flawed in this respect, cultural differences will most likely manifest in differing intensities (strongly agree vs. agree), and thus will not skew the results dramatically. Finally, it is important to note that differences in responses *do* reflect differences in the bilingual experience, and thus rating two culturally different stories as less similar does reflect an actual difference in experiences. An extensive discussion can be found [2].

4.2 Quantitative Statements

Certain aspects of the multilingual experience are more quantitative. For instance, each speaker considers themselves knowledgeable in some number of languages, and each speaker uses, on average, some number of languages each day or week. The quantitative statements provide further grounds for comparison between stories by focusing on the more objective aspects of a bilingual's experience.

In order to facilitate accurate matching, it is important to have some information that overlaps for all stories. To this end, all quantitative questions are required; a recount that does not answer them is, in my opinion, incomplete, and thus invalid for comparison. The quantitative questions provide relatively basic data, such as the number of languages spoken, the use cases of those languages, and the existence of a single native language. While high confidence is not required, all questions must be answered at some level of confidence. [3]

5 Comparison Implementation

In order to determine the magnitude of difference between two stories, I will adapt a method known as *Cosine Similarity* . Hold up your hand, and look at the angle between your index and middle finger. The idea behind cosine similarity is that the fingers are most similar when the angle between them is close to zero. So, when your index and middle finger are touching, they are most similar; when one points down and the other points up, they are least similar.

More mathematically, each story corresponds to a vector in many dimensions. Stories that are most similar are stories whose arrows are nearly parallel and have a small angle between them. The Cosine Similarity is then computed as follows [3]:

$$\phi = \frac{A \cdot B}{||A|| * ||B||}$$

Similarity varies between -1 and 1. Stories that are most similar have a ϕ value that is closer to 1. Stories that express opposing views have a ϕ value closer to -1. Stories with a ϕ value of 0 are different, without opposite views.

Cosine similarity has the benefit of providing a metric of similarity that is independent of the number of characteristics compared. This ensures that, in the case where a question must be excluded from the comparison because the question is not answered at a sufficient

[3] See Section 11.1 for more information.

confidence level by one or both stories, the measurement of similarity is still valid and can be compared to the measurement of similarity between other stories. This flexibility is crucial, as the number of aspects on which any two stories are compared will likely differ in each comparison. [4]

6 Handling Gaps in Data

Addressing data gaps is a well known and relevant problem in data analysis. Real data is often missing some components. In addressing the issue of analyzing data despite missing pieces, two basic approaches are used: deletion and imputation [5]. Neither of these techniques is sufficient alone for handling missing data in this case. To this end, quite sophisticated imputation methods have been developed; these tools are overly complex for our applications here [1].

Deletion refers to eliminating any data that is incomplete [5]. In the context of this project, deletion would mean that, if story A and B are being compared, the comparison would be based solely on questions that both A and B both answered explicitly. While similarity is based on definitive responses, this approach has clear drawbacks. In particular, much of the bilingual experience is implied, not explicitly stated. If deletion is used, then, since many of the opinions and experiences of a bilingual are conveyed without explicit answers to questions, much of the present, implicit data will be excluded from the comparison, in favor of the small subset that is explicitly addressed. Resulting comparisons could be based on a small number of questions, potentially producing false similarity, if two stories overlapped on a few questions quite closely.

Imputation refers to estimating missing data based on other responses [5]. There are many different possibilities for generating such an estimation. In this context, if A did not answer a question, imputation might mean using the answers of other stories (either all or some subset), to construct an average response that predicts what A might respond. There are clear drawbacks here as well, mainly stemming from the complexity of any given person's bilingual experience. Experiences can vary dramatically, and there is little guarantee that the answers of one person can accurately predict the answers of another. Further, failure to answer a question can suggest multiple interpretations. For instance, if A does not mention negative experiences with bilingualism, it could suggest that A did not have

[4] In order to make the reported similarity ranking intuitive, I choose to convert the Cosine similarity into a percentage before displaying the result to the user. This decision is discussed in Section 11.3.

any negative experiences. However, it is possible that A did have negative experiences with bilingualism, but felt uncomfortable sharing those personal experiences with others. Blindly imputing default or average values to estimate A's plausible response erroneously simplifies A's experiences and could produce inaccurate results.

6.1 Data "Between the Lines" and Multiple Confidence Levels

Above, I emphasized that neither deletion nor imputation alone provide a satisfactory solution to address gaps in bilingual experience data. Note however, that deletion and imputation represent differing ends of a continuum for accounting for missing data. Suppose that deletion was used. Then, stories marked as similar would likely share a few, salient features. However, more subtle aspects of the stories might differ, and these aspects may not be accounted for in the metric of similarity. Suppose that imputation was used. Then, stories marked as similar would probably share more subtle characteristics that may be implied, but not stated, even as more salient differences could exist.

Above, I remarked that each qualitative response had an associated level of confidence. This level of confidence is used to implement the appropriate comparison based on the user's desired levels of imputation and deletion. Different ends of this continuum are appropriate in different contexts. Therefore, it will be up to the user to determine, at the beginning of the analysis, to what extent deletion vs. imputation should be used for handling gaps in the data. The user does this by selecting a **Minimum Confidence Threshold**. *When comparing stories A and B, only questions that both A and B answer with confidence of at least the selected threshold will be used in analysis.* [5]

Thus, if the user wants comparisons to be made on salient, definitive aspects of the bilingual experience, stories will only be compared based on questions where both answers have a high level of confidence. This approach effectively limits the amount of imputation used. By contrast, if a user would prefer to compare on both salient and more subtle aspects, then they can pick a low minimum confidence threshold, and stories will be compared based on a broader range of questions.

The use of confidence levels effectively allows the user to choose the balance between imputation and deletion appropriate for their needs.

[5] This algorithm was not the first idea for determining similarity. A discussion of other, previous attempts can be found in Section 11.2.

6.2 Quantifying the Corpus

Above, I established both the types of questions and statements I will use to quantify each story and the necessity of pairing each response with a confidence level. Below, in Section 10, I have also specified the questions I will use.

For each story, I will read though and attempt to make a judgment of how the subject might respond to each question in my set. In some cases, questions will be answered explicitly, in which case I will be able to report an answer with high confidence. In other cases, my estimation will perhaps rely on more assumptions or implications. In these cases, I will mark my answers with lower confidence levels, and they will only be used when the confidence of my prediction exceeds the minimum threshold established by the user. [6] In the case where I cannot estimate the answer to a question with even low confidence, I would not like to make comparisons based on that question, as there is simply not enough information to make an accurate judgment. This means that if A does not provide an answer to a question, the unanswered question will never be used to compare A to other stories.

An important issue arises when determining how omitted information should be interpreted. For instance, as discussed above, if A's recount does not mention negative experiences due to bilingualism, there are two drastically different interpretations: A perhaps did not have any negative experiences, or, alternatively, did not wish to share negative experiences. It is impossible for me to know about elements of A's experience that were present but not shared. Therefore, when estimating responses and confidence levels, I will assume that no information is being suppressed, and thus would conclude the former possibility. I recognize that this assumption may cause some false matches, but believe it is unavoidable. I can only use the given data, and compare based on what is actually indicated by the recount.

Finally, it is important to recognize some inevitable variation in the corpus resulting from differences between interviews. Some topics simply may not have been addressed due to time constraints or other reasons. Further, topics that were addressed may have been framed differently from one interview to the next. A more thorough discussion of some potential sources of variation can be found in Section 13.3.

[6] It is important to recognize that my estimations are inherently subjective. While confidence values are fairly effective at addressing this, avoiding subjectivity altogether is impossible. A more detailed discussion can be found in Section 13.2.

7 Concise Description of Algorithm

- Select a story, A, that we want to compare to all other stories.
- Get choice from user about desired confidence level
- Ensure that A has answered all required questions.
- For each story B in the corpus:

  For each question answered by A with acceptable confidence

  If that question is answered by B with acceptable confidence

  Include that question in data for analysis

  Otherwise

  Exclude that question from analysis

  Having constructed a subset of the data with appropriate confidence, compute the Cosine Similarity, ϕ, between the two stories.
- Convert the Cosine Similarity into a percentage value.
- Report all stories compared to A in order of decreasing percentage similarity.

8 Program Control Flow

The following section describes the main functions of the program and different prompts you may encounter.

1. The program looks for a stories folder and parses all files in that folder. Any errors are printed to the screen.
2. The program prompts the user to ask if they want to incorporate a new story (such as their own) into the corpus for this session.

   If Yes:

   The program asks a series of questions, generating a story classification file, and parses their story file. If the user indicates they are entering their own story, confidence values are assumed to be "High Confidence." Otherwise, each question is prompted with a confidence value.

3. Program Main Loop: User selects an option for what they want to query the corpus for.

4. When the user wants to exit:

   If they have input their own story, ask them if they want to submit it to the corpus

5. Exit the program

Compare Two Stories Side By Side :

1. Select desired confidence threshold
2. Choose a first story from the list of stories in the corpus
3. Choose a second story from the remaining list of stories in the corpus
4. Characteristics are listed in tables, depending on how distant the characteristics are.

Compare One Story to All Other Stories :

1. Select desired confidence threshold
2. Choose a first story from the list of stories in the corpus
3. Cosine similarity is calculated between the selected story and all other stories, the similarity is converted to a percentage, and stories are reported in decreasing order of similarity. The number of characteristics used in the comparison is also reported, as this could differ between pairs of stories depending on which questions were answered and the confidence of those answers.

View A Single Story File :

1. Choose a story from the list of stories in the corpus
2. The responses to all questions and the confidence of each response are printed in a tabular format.

Filter Stories by Attribute :

1. Choose a question of interest
2. Select which responses to include

3. The program outputs a list of the stories that responded to the selected question with the selected response, along with the confidence level of that response.

View Corpus Statistics :

View Distribution of Responses shows, for each qualitative question, the number of responses of each type at each confidence level.

View Distribution of Confidence Values shows, for each qualitative question, the number of responses at each confidence level for that question (i.e.: how many recounts have an answer with low, medium, or high confidence).

View Mapping of Question Codes shows the question in a readable format, along with the code used in the story recount file to code that question.

9 Running the Program:

The following instructions are intended for a Mac.

1. In the top right corner of the screen, click the magnifying glass, and type `Terminal`
2. You should see a square box show up on the screen with a prompt that ends in a `$`.
3. Type in the following command: `ssh ling001@cs.swarthmore.edu`
4. It will prompt you for the password: **********
5. The program will begin running automatically, following the control flow indicated above.

10 Questions

Each question is classified by the category of the bilingual experience it intends to address. As noted above, the chosen categories include language acquisition, experience, use, identity, and intention. A discussion development of these questions and statements and some potential questions to add in future analysis can be found in Sections 11.1 and 13.1.1 respectively.

Quantitative Questions

Acquisition: How many languages do you speak?

Answers: 1, 2, 3, 4+

Acquisition: How many languages were you exposed to before beginning school?

Answers: 1, 2, 3, 4+

Acquisition: How many languages are you fluent in?

Answers: 1, 2, 3, 4+

Use: How many languages do you use each day?

Answers: 1, 2, 3, 4+

Use: How many languages do you use each week? [7]

Answers: 1, 2, 3, 4+

Use: How many languages do you use each year? [7]

Answers: 1, 2, 3, 4+

Identity: I have a single native language that I feel closest to.

Answers: Yes, No

Qualitative Statements

- Acquisition: The language I spoke in primary school is now my dominant language.

  For the purposes of this question, primary school is generally considered to be elementary age. If multiple languages were spoken during that period (due to moving, for instance), or the speaker does not indicate a single dominant language, this question may be unanswered or neutral with low confidence.

- Use: One of my languages is used only with a specific group of people such as family or a set of friends.

  For the purposes of this question, people, not experiences, are emphasized. This question determines if there is a language that is only used for maintaining contact with a small group of people, such as an old friend or elderly family members. It does not address whether a particular language is restricted to a situation or context, such as international travel.

[7] In actuality, the relative number of languages used (i.e.: the difference between the number used daily, weekly, and yearly periods) is stored by the program. This design decision is discussed more in Section 11.1.

- Acquisition: I learned one of my languages primarily in a language class (and not through exposure in society).

  This question determines whether language acquisition has been primarily due to the surrounding context, or whether the speaker has taken a language class. This can have implications for the speaker's future intentions and desires regarding language. Additionally, it could be interesting to analyze whether speakers ever consider themselves fluent in languages learned through classes, rather than context.

- Intention: I wanted to be bilingual and learn the languages I've learned.

  This question alludes to the speaker's initial opinions of learning multiple languages. In some cases, the speaker is neutral about learning the language, having learned it out of necessity.[8] In other cases, the speaker may have hated their grandparent's compulsory language lessons. [9]

- Intention: I want to learn more languages or improve my existing language skills.

  This question indicates existing attitudes towards bilingualism. In some cases, it presents an interesting contrast with the question above, representing a shift in opinion over time with regard to bilingualism. [10]

- Use: The first language I learned is used when I have strong emotions

  This question indicates what language is chosen when the speaker experiences strong emotions, such as happiness or sadness, and thus may feel the need to communicate quickly and expressively. [11] It also indirectly indicates to what extent the first learned language is still used and / or accessible to the speaker.

- Use: I choose to address people in the language that I think will make them most comfortable.

  This question indicates the speaker's langauge choice. Some speakers indicate that they choose to use language that will make others feel comfortable, [12] while others indicate

[8]See *Native to Nowhere: Travel Multilingualism and Identity* (Poyer); *Identity Through the Bilingual Experience* (Barrientos).

[9]See *The Bilingual Experience: A Fusion of Language and Culture* (Nasseri).

[10]See *The Bilingual Experience: A Fusion of Language and Culture* (Nasseri); *Interview with Evangelos C.* (Molloy).

[11]See *Oui Oui Croissant* (Quevedo); *Excuse My French: A Bilingual Linguistic History* (Stigliani).

[12]See *A Glimpse into La Vida de Tyler Welsh* (Carney); *Bilingual Interview* (Conca).

that they may intentionally avoid using a language in order to separate themselves from others. [13]

- Use: I feel frustrated by words or expressions in one language that are not in another.

  This question indicates whether the speaker feels constrained by use of a single language. In some cases, speakers indicate great frustration when certain ideas do not easily translate, [14] while for others, this frustration does not seem to be as prevalent.

- Identity: Being bilingual is an important part of who I am.

  This question references the extent to which bilingualism is a defining characteristic of a speaker's identity. How might the speaker feel if they were not bilingual? To what extent would they consider themselves a different person? This question is independent of the one below, as it does not address whether the speaker enjoys having this part of their identity.

- Experience: I enjoy being bilingual.

  This question indicates the speaker's current attitude towards bilingualism. Is it fun to be a bilingual? Useful? Frustrating since others can have difficulty understanding them? Note that bilingualism may be an important part of a speaker's identity without them enjoying their bilingualism.

- Identity: People call me different names depending on the language used.

  This question indicates whether the speaker uses multiple names. Some speakers, due to cultural or phonological reasons, choose to adopt a different name in each language. [15] Others insist on using a single name, even if it is mispronounced. [16]

- Experience: Overall, people judge me positively when they realize I am bilingual.

  This question indicates the response of others (peers, family, and society at large) to encountering the bilingual speaker. In some cases, bilingualism is greatly encouraged;

[13] See *Interview with William Lin* (Noyes); *Native to Nowhere: Travel Multilingualism and Identity* (Poyer).
[14] See *Interview with French and English Bilingual Individual* (Barnett); *Cristina's Linguistic Journey* (Kazaklar).
[15] See *Mot Couc Phong Van Voi Sally* (Cheney); *Mixing Chinese and English - Interview with Pearcela Geng* (Wang).
[16] See *Identity Through the Bilingual Experience* (Barrientos).

[17] in others, bilingualism may lead to negative associations with other cultures or nations. [18]

- Intention: I want / wanted my children to be bilingual.

  This question indicates the speaker's current view of bilingualism insofar as the extent to which bilingualism is something that would like to pass on to their children.

- Experience: I am still viewed as a native speaker in the language I learned first.

  This question indicates the extent to which the first language learned is retained by determining how other native speakers view the speaker. In some cases, the fluency obtained originally wanes over time, [19] while in other cases the speaker remains fluent. [20]

- Experience: I find that my relationships with other bilinguals are deeper than those that involve a single language.

  This question indicates whether the speaker has bilingual relationships, and whether those relationships differ from monolingual relationships. In some cases, speakers indicate that their relationships with other bilinguals are deeper and more meaningful due to the increased ease of communication between them. [21]

- Experience: Knowing multiple languages has given be professional or tangible benefits.

  This question indicates whether the speaker derives tangible benefits - such as scholarships, increased job opportunities, or prizes - due to their bilingual knowledge. Some speakers emphasize that bilingualism has been useful to them in this regard, [22] while others make no mention of tangible benefits, focusing on other topics, such as the personal benefits of feeling connected to their past. [23]

- Experience: Bilingualism can have negative effects.

[17] See *Multilingualism: The Key to Multiculturalism* (Ehsani); *Interview with a Bilingual Individual* (Erskine).

[18] See *Being Bilingual* (Lucas); *Oui Oui Croissant* (Quevedo).

[19] See *Mot Couc Phong Van Voi Sally* (Cheney); *Translating Identities: Bilingualism and Cultural Identity* (Senft).

[20] See *Interview on Bilingualism with Glen Rico* (Rico).

[21] See *Identity Through the Bilingual Experience* (Barrientos); *Interview with French and English Bilingual Individual* (Barnett).

[22] See *Oui Oui Croissant* (Quevedo); *Mixing Chinese and English - Interview with Pearcela Geng* (Wang).

[23] See *The Bilingual Experience: A Fusion of Language and Culture* (Nasseri).

This question indicates the extent to which the speaker has seen or been exposed to negative judgment on account of their bilingualism. It is intentionally indirect to avoid triggering upsetting memories in the speaker. Rather than just asking if the speaker has had negative experiences, it attempts to estimate this by determining to what extent the speaker is aware of negative consequences that may result from being bilingual. Some speakers indicate that they have had or seen negative consequences from their bilingualism, [24] while others report that there are no cons to being bilingual. [25]

11 Classification Reflection: An Evolving Method

Determining the correct algorithm and questions to use when classifying the corpus was a challenging, thought-provoking process. In this section, I describe some of the changes made from my original intentions, and the rationale behind those changes. The result is, in my opinion, a reasonable algorithm for quantifying stories in the corpus that addresses the major components of the bilingual experience while allowing for gaps in the stories.

11.1 Changes to Questions

While reading the corpus, I made a few changes to the questions I was using to classify the corpus.

Intentionally Excluding Specific References to Languages : I made a conscious decision not to include any information about specific languages in my statements. This reflects my opinion that the bilingual experience transcends any single language or group of languages. Therefore, my questions center on the experiences of the speakers, without focusing on the particular languages spoken.

Quantitative Questions Include Confidence : Initially, I thought that the quantitative questions about language knowledge, fluency, and use were sufficiently integral to the bilingual experience that I would be able to answer them with high confidence for all stories. I quickly realized that this assumption was incorrect; the quantitative questions, especially those regarding relative frequency of language use, while often answered, were often not answered with high confidence.

[24] See *Growing Up Bilingual* (Dou).
[25] See *Mot Couc Phong Van Voi Sally* (Cheney).

I still feel that the aspects of the bilingual experience referenced by the quantitative questions, such as the number of languages known and used, are defining characteristics of the bilingual experience, and thus feel it is appropriate to make these questions required. Thus, the program will not accept a recount without answers to these questions. However, I have relaxed the constraint that these questions be answered with high confidence. In many cases, I felt like a high confidence rating for these questions would artificially overstate my own confidence in the answers, thus leading stories to erroneously rank similarly solely on implied language use answers.

Quantitative Usage Questions Compare Relative Frequency : Initially, I intended to have three questions about language use, indicating the number of languages used each day, week, and year. Further, I intended to compare directly on these answers. However, I realized in reading the corpus that the relative frequency of use, rather than the number of use itself, was often more informative. For instance, I found that most speakers used all languages at least once a year, which meant that the yearly question was redundant with the number of languages spoken in total.

To remedy this, I instead compare on the *relative* number of languages used daily, weekly, and yearly. This means that, if the user reports speaking 4 languages each year and 3 languages each week, then the comparison on the question of yearly language use will be done on $4 - 3 = 1$, indicating that the speaker uses one language on a yearly, rather than weekly, basis.

However, I feel like the relative question is more subtle, and likely more counter intuitive, than the original question of the number of languages used in a given period. For that reason, the program (specifically, the part that reads in a new story and generates the story file) does the computation behind the scenes, transparently to the user. When inputting a story, the user indicates the number of languages used in each period, but the relative use is what is stored and used in subsequent comparisons.

Rephrasing the Negative Experiences Question : I phrased the question about negative bilingual experiences carefully: "Bilingualism can have negative effects." This phrasing resulted from an early observation that few stories explicitly referenced negative judgment resulting from bilingualism. In most cases, negative experiences were subtly implied instead.

There are a number of reasons this might be the case. Negative experiences are often raw memories of painful judgment and feelings of exclusion. It is possible that, when

interviewed, an interviewee did not share information of this nature, simply because they felt uncomfortable sharing such information with a class of strangers. Alternately, they may not have wanted to relive the painful experience by telling it to others.

The negative experiences question is carefully crafted. First, it is intended to be indirect, hoping to avoid triggering hurtful memories. At the same time, it is slightly broader than a question about personal negative experiences. This allowed me to appropriately note and quantify the vague references tp negative experiences, without necessarily needing to know or include personal negative experiences in the recount.

Separating Prior and Future Goals for Bilingualism : Initially, my intention was to have a single question: "I want / wanted to learn multiple languages." Early on, however, I realized that this single question was conflating two very distinct issues. Some speakers [26] indicate a strong desire to learn or improve their language skills, even as they did not initially want to learn multiple languages. I wanted to be able to capture a shift from past intentions to future intentions. Therefore, I separated the single question into two questions, one to address past views of bilingualism, and the other to address future goals for bilingualism:

1. I wanted to be bilingual and learn the languages I've learned.
2. I want to learn more languages or improve my existing language skills.

This separation turned out to be instructive. While five stories were ranked as disagreeing with the first statement at medium or high confidence, none were ranked as disagreeing with the second statement at similar confidence levels. [27]

11.2 Before There Were Confidence Values: Alternate Approaches Considered

The first issue that arose quickly was that some of my questions were not answered by all recounts. In this case, what should be done about comparisons? My first instinct was to assign a "default value" to each question, which would be used if two stories were being compared, and one had answered a question and another had not. The default value would,

[26] See *The Bilingual Experience: a Fusion of Language and Culture* (Nasseri); *Interview with Evangelos C.* (Molloy).

[27] Full disclosure necessitates mentioning that a single story is listed, at low confidence, as disagreeing with the second statement.

hopefully, provide a reasonable estimate for what the other speaker was likely to respond, thus facilitating accurate and complete comparisons.

I assigned each question a default value, but quickly realized after reading through some of the corpus that my assigned values were overly simplistic. Many stories did not answer questions explicitly, but the default value was different from what I believed was probably the most likely answer by that speaker. For instance, few speakers explicitly were asked about whether they believed bilingualism could have negative effects, but many alluded to bilingualism as overwhelmingly positive or referenced times in their life where bilingualism had painfully set them apart. I needed a method to encapsulate data that was not explicitly provided, while also providing a mechanism to treat such inferred data differently from explicitly provided data.

Adding confidence values neatly solves this problem of default values and unanswered questions. When a question was unanswered, I estimated a response and also noted how confident I was in that response. This allows the user to differentiate between inferred and explicitly provided data, while also providing the flexibility to incorporate such inferred data into the database.

11.3 Reporting Similarity as a Percentage

I decided not to report the Cosine similarity directly to the user, instead opting to rescale the similarity to be between 0 and 100. This has the benefit of being less confusing for the user, since they will be able to readily use their intuitions with percentages to interpret similarity. Since it is unlikely that any particular user will closely read this implementation paper and understand that the range of similarity goes from -1 to 1, many might erroneously assume that the scale is only positive. This is especially plausible, given that the stories currently in the corpus tend to be rather similar, and the results are often between 0.5 and 0.95. If I reported cosine similarity directly, users might erroneously assume that 0 indicates least similarity, when in actuality -1 indicates this. Rescaling to positive values between 0 and 100 will hopefully facilitate more accurate, informed analysis without requiring users to understand the details of the implementation.

12 Extensions and Further Work

I hope that this project will have continued value beyond this semester. To that end, I have attempted to make the software and design extensible. The questions on which each story

is quantified are intended to apply to many stories, and thus provide a common basis for comparison. Further, the software is designed to make adding new stories to the comparison simple: incorporating an additional story consists of quantifying the story as previously described and placing the information file in a folder with all other stories to be used.

There are many possible extensions that could further enhance this project:

Additional Data : Currently, the database contains only the stories from Introduction to Linguistics students this semester. It would be exciting to see more stories added to this database to facilitate better matches and more interesting comparisons. Such data could either come from additional semesters of students, or be crowd sourced online to a much broader audience.

Selective Comparisons : Currently, the program compares stories based on all available data. However, in certain cases, it may be desirable to only compare stories based on a few, selected characteristics. One relatively simple extension could allow the user to selectively exclude certain characteristics from the algorithm, providing for a more focused comparison.

Graphical / Online Interface : Currently, the interface is quite simple and text based. In order to make this program easier to use, it would be helpful to provide a clickable graphical interface. Further, adapting this program for the Internet would expand access to millions of people, potentially allowing for greater utility of the program. While I do not have permission to share this semester's stories beyond the class, in subsequent semesters, perhaps some individuals will consent to have their story posted online, which could facilitate making this goal of broader access a reality.

Similarity Feedback : Many ratings systems used to predict similarities, such as the Netflix or Amazon matching algorithm, use subsequent ratings or actions to further improve prediction algorithms. In this context, a rating system might allow the user to input their own judgment on the similarity of two stories, thus providing supplemental insight to the quantified data. Such feedback would be especially useful for judging similarity on less explicit characteristics of a story, and could allow for higher confidence in judgments extrapolated from the stated recount. It would be exciting to see feedback incorporated into this program to allow all users to collectively contribute towards an accurate matching system.

13 Retrospective Reflection

Looking back, a few aspects of my project, and the corpus overall, seem especially noteworthy.

13.1 Validity of Criteria Used for Classification

In the process of classifying bilingual stories, I gained confidence in the general validity of the criteria used for classification. In most cases, answering all questions required a thorough reading of the recount; very little information was included that was not referenced, in some way, by the questions. That said, there are a few places where my questions did not encompass elements of the bilingual experience, and one question that was rarely answered.

There are two aspects of the bilingual experience that were referenced in more than one story but not encompassed by my criteria. A few stories talked explicitly about language use and poetry; [28] others referenced language use when intoxicated. [29] However, the majority of recounts did not reference these recounts. While I feel that one could certainly include either one as an element of the bilingual experience, I don't feel it is necessary, as the sentiments expressed by both elements are referenced, somewhat, by the questions "Being bilingual is an important part of who I am" and "The first language I learned is used when I have strong emotions." This is because, if poetry is an important aspect of a speaker's life, and poetry and bilingualism are intertwined, then the sentiment of bilingualism as it relates to poetry can be encompassed in the broader question of how bilingualism relates to identity. In a similar way, the sentiment referenced by language use when intoxicated is similar to the sentiment referenced by questions about language use and strong emotions. Therefore, even though these aspects are not explicitly included in my criteria, I believe the information conveyed can still be retained using my criteria.

The vast majority of recounts did not address whether bilingual relationships were deeper than relationships using a single language. Nevertheless, I feel that this remains an important characteristic of the bilingual experience. The speakers who mentioned a sense of connectedness with other bilinguals talked passionately about how these relationships allowed for easier, more fluid communication. Further, I believe that other bilinguals would likely agree with this statement, had they been asked directly. [30] Therefore, even though few people responded to this question, I still feel it is an important aspect of the bilingual experience,

[28] See *The Polyglot of Wharton AB 1st* (Tarlin); *The Poet* (Miller)

[29] See *The Polyglot of Wharton AB 1st* (Tarlin)

[30] Anecdotally, my interviewee mentioned that she shared three languages with her husband, and would have likely agreed with the question, had I asked more generally about relationships with other bilinguals.

and I would not remove it.

13.1.1 Questions to Add

While I feel that my questions to a reasonable job of incorporating many aspects of the bilingual experience, there are some statements that I would add to the list used to quantify stories.

I would add one additional question about language acquisition. Above, I talked at length about how I split the question of attitudes towards bilingualism into two questions - one about prior attitudes and the other encapsulating future goals. It might be interesting to add an additional question: "I had no choice in becoming a bilingual." With the first questions above, those forced to learn languages typically fall in the category of neutral or disagreement responses, alongside those who passively picked up language without strong emotions either way. It is likely that these situations are different, and would be interesting to see what sort of analysis could be done analyzing this difference more closely.

I would also modify the question about language choice. Language choice can be motivated by at least two distinct factors. Some recounts emphasized that the comfort of others was an important factor, noting that they would speak a language used by others around them. However, in other cases, langauge choice seemed to be motived by a desire to fit in or stand out relative to others, without concern about comfort. The existing question conflates these motivations to some extent, so I would modify and add questions to separate these two motivations.

13.2 One Opinion - Inherent Subjectivity of Estimation

When creating the questions, I tried to make questions that would be reasonably clear cut and objective. I figured that the aspects I was measuring were relatively universal, and that, surely, there would be a consistent, clear answer for most, if not all, questions.

For a few questions, this turned out to be true. For instance, most recounts explicitly answered whether the bilingual subject wanted their kids to be bilingual as well, and nearly all answered whether the language associated with education became the dominant language. [31]

[31] One fascinating element of the corpus was that, for many speakers, the dominant language was the language used in school. This is likely a reflection on the fact that peer influence typically is stronger than parent influence when it comes to language, and people will thus generally adopt the language of their peers in school.

I quickly realized that, despite my best intentions, the questions did not always have clear cut answers. Probably the clearest example of a question that was subtly implied, but not explicitly addressed, was the question about negative experiences resulting from bilingualism. As noted above, most recounts made vague references to times where they had been judged for being bilingual, but few actually stated that bilingualism could have negative effects. Since negative experiences are often alluded to, but not explicitly discussed, there were often no clear answers, and very few stories were listed as responding to this question with high confidence.

Something as simple as a first language learned turned out to be more complex than anticipated. One of the questions - " I am viewed as a native speaker in the language I learned first" - presumes that there is such a language. While one approach is to simply not answer the question if there is no clear first language, there are still instances of ambiguity in some cases. [32]

The question about language use, "I choose to address people in the language that I think will make them most comfortable," was also less objective that I expected. To my surprise, most recounts did not explicitly address how the language used was chosen, and, of those that did, many did not explicitly address comfort as a factor used in determining language choice. In many cases, I inferred the extent to which the comfort of others was a factor in language choice from the situations in which a language was said to be used. For instance, if a bilingual said that Korean was only used after hearing others speaking Korean, I inferred, with moderate or high confidence, that comfort of others was a factor.

All of these cases are illustrative of the fact that some of the quantification, intended to be definitive and objective, has some subjective elements. Many questions are not definitively answered, and inferring how a speaker would respond is difficult. I worked quite hard to attempt to rank each story as objectively as possible, however, stories are complex, and there are almost certainly some elements others might judge differently. People with different backgrounds from mine may focus on different characteristics and, in doing so, come to differing conclusions about the bilingual's experience.

[32] Consider *Being Bilingual* (Lucas), in which a bilingual speaker moved at age two from the US to Iran. Should English be considered the first language learned in this case? Or should the first language learned be the first language in which fluency is achieved?

13.3 Inevitable Variation Between Interviews

I think it is important to recognize the differences between interviews and stories used to form the corpus. Some of the variation between the stories told may reflect differences in the questions asked, specific wording of those questions, and the context of the interview. These differences may artificially differentiate stories that, in actuality, are similar.

Different interviewees are asked different questions: Some questions, such as the question about relationships with other bilinguals, were not addressed in many recounts. However, as I alluded to above, I believe that most bilinguals would agree with a statement about relationships with other bilinguals being deeper than those with other monolinguals. Therefore, I believe that some of the variation within the corpus, especially with questions that were unanswered, may be more of a reflection of questions that were not asked, rather than statements about the bilingual experience.

Relationships between the interviewer and interviewee differ: The nature of the interviewer and interviewee could influence both the questions asked by the interviewer and the level of detail of responses given by the interviewee. For instance, in the case of questions about negative experiences or future intentions with children, an interviewer simply may not feel comfortable asking such an intimate question of their interviewee. Other times questions may not seem relevant or appropriate. For instance, it may not seem appropriate to ask a hall mate about professional benefits from their bilingualism, under the assumption that the hall mate has not held any professional jobs. Similarly, if the interviewee has not ever mentioned using multiple names, an interviewer may not ask explicitly, assuming that such a question is irrelevant to the interviewee.

The same question can be asked and interpreted in different ways: It's important to recognize the strong effects wording of a question can have on the response. For instance, a speaker might respond differently to the questions "Is bilingualism an important part of who you are" and "How would you feel if, suddenly, you were no longer bilingual," even though both questions address bilingualism and identity. Similarly, asking if speakers if a particular language is associated with strong emotions could yield a different response than if the same person was asked if a particular language was associated with a single emotion, such as anger. Presumably, in the latter case, the answers would likely be more definitive, as the scope of the question is narrower. Further, interpretation of questions may differ between speakers. For instance, one interviewee might interpret "professional benefits" to include admission to college, while another might exclude such academic benefits.

Therefore, it's important to recognize that there is some inherent variability in the corpus

due to which questions were asked, and the wording of those questions. This variability is unavoidable however, given that the interviews were conducted with separate bilinguals by separate interviewers. Optimally, interviewees would answer the questions directly, but even in this case some variation between responses remains inevitable.

13.4 Relative Demographic Consistency of the Corpus

One thing that struck me in reading the corpus is that many of the stories have a lot of demographic similarity in terms of age and education backgrounds. It is important to note that those chosen to participate as interviewees may not be representative of bilingual speakers as a whole.

The vast majority interviewed expressed a desire to improve language skills. At a medium level of confidence, over half of the stories interviewed expressed some desire to learn more languages. This is likely a reflection of the fact that the corpus includes a disproportionate number of bilinguals of college age, many at Swarthmore. Students in the academic environment fostered by college are probably more likely to be more driven to improve their language skills than other bilinguals.

The majority interviewed expressed, with strong confidence, that bilingualism was an important part of their identity. This is likely a reflection of the fact that students, when choosing others to interview about bilingualism, are likely to choose others that are open and approachable when asked about their bilingualism. The people interviewed are likely to be individuals whose bilingualism is shared and salient, since these will be the people known by the interviewers to be bilinguals. As a result, those interviewed probably feel bilingualism is more important to their identity than the average bilingual.

Most stories expressed a desire to raise children as bilinguals. This is likely linked to the item above. Individuals who do have *not* had positive experiences as bilinguals are also unlikely to share their bilingual identity with others, including their children and peers, and thus are unlikely to be selected or consent to a bilingual interview. Thus, the corpus overestimates the proportion of bilinguals who want to raise their children as bilinguals.

14 Conclusion

Writing this program, classifying the stories, and using the results has been a thought provoking, challenging experience for me. However, I feel that I have been successful based on the creation of the following deliverables:

- A similarity measurement system that handles gaps in data in a reasonable and flexible way.
- A set of questions whose answers, in my opinion, touch on a substantial portion of the bilingual experience as a whole.
- A useful, extensible program tool that easily allows other linguists or bilinguals to input their own data, and analyze trends and compare similarities between other stories

15 Evaluation Criteria

I will consider myself successful in this project if I have produced the following deliverables:

1. A reasonable, justified set of characteristics that make up the bilingual experience.
2. A reasonable, justified approach to classifying stories incorporating those characteristics.
3. Classifications of stories in the corpus according to the criteria developed above.
4. A reasonable, justified algorithm for ranking similarity between stories using the classification described above, which handling cases where stories do not address certain characteristics.
5. A functional, extensible program that:

   Implements the described similarity maching algorithm

   Allows for easy extraction and use of data in the corpus

   Allows new stories to be added into the database by the user.

Further, I will consider myself successful if my deliverables reflect the overall goals of the final project by:

1. Showing deep engagement with the existing corpus materials.
2. Demonstrating critical thinking regarding both the criteria for classifying stories and the ranking algorithm resulting from such engagement.
3. Facilitating connections between stories in a simple way.

References

[1] Alan C Acock. Working with missing values. *Journal of Marriage and Family*, 67(4):1012–1028, 2005.

[2] Steven J Heine, Darrin R Lehman, Kaiping Peng, and Joe Greenholtz. What's wrong with cross-cultural comparisons of subjective likert scales?: The reference-group effect. *Journal of personality and social psychology*, 82(6):903, 2002.

[3] Christian S. Perone. Cosine similarity. 2014.

[4] John S. Uebersax. Likert scales: Dispelling the confusion. 2006.

[5] Richard Williams. Missing data part 1: Overview, traditional methods. March 2013.