PHONOLOGICAL MARKEDNESS EFFECTS ON NOUN-ADJECTIVE ORDERING

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Katherine Marie Blake

August 2022

PHONOLOGICAL MARKEDNESS EFFECTS ON NOUN-ADJECTIVE ORDERING

Katherine Marie Blake, Ph.D.

Cornell University 2022

Traditional theories of grammar posit that sentence formation begins with the underlying hierarchical structure generated by the syntactic component, which is then given meaning and sound interpretations by the semantic and phonological components (Chomsky, 1965). This dissertation provides evidence that word ordering is conditioned by the avoidance of phonologically-marked structures, challenging strictly feed-forward theories of grammar.

Previous work in various languages shows phonological effects on word ordering, such as: binomials and sentence formation in English (Morgan, 2016; Breiss and Hayes, 2020); compounds in Navajo and English (Martin, 2011); topicalization in Serbo-Croatian (Inkelas and Zec, 1995); and noun-adjective ordering in Tagalog (Shih and Zuraw, 2017). This work examines the effects of six phonological markedness constraints on flexible noun-adjective ordering in five languages based on the analysis of large speech corpora. Constraints include stress clash and lapse, vowel hiatus, consonant clusters that disagree in voicing, consonant clusters that agree in place of articulation, and relative word length. The languages analyzed are French and Italian (Romance), Polish (Slavic), Hindi (Indo-Aryan), and Modern Standard Arabic (West Semitic).

The first of two central hypotheses of this dissertation is that only those phonologically-marked phenomena that are avoided with phonological repairs will also be avoided via word ordering. This is tested by looking at the avoidance

of two types of phonological constraints in each language, those that are active in the language (e.g., vowel hiatus in French), and those that are inactive in the language (e.g., consonant clusters at the same place of articulation in French). In general, results support this hypothesis. For example, vowel hiatus was found to be avoided via word ordering in French noun-adjective pairs, but not in Italian where it is not active; and, there is evidence that consonant clusters across the word boundary that share a place of articulation are not avoided in French, but are avoided in Polish.

The second central hypothesis of this work is that phonological effects on ordering are stronger if semantic differences between orders are minimal. To test this hypothesis, semantic difference between the prenominal and postnominal form of an adjective is quantified by the similarity between positional embeddings; noun-adjective pairs with an adjective above a determined similarity threshold are analyzed separately from those below the threshold. Results did not support this hypothesis. There were no phonological effects that were greater or present only in the semantically-similar dataset.

Overall, this dissertation argues that noun-adjective ordering is conditioned by phonological markedness effects at the prosodic, syllabic, and segmental levels. This finding is discussed with respect to existing theories of the interface between phonology and syntax, concluding that at a minimum, phonological factors compete with others to determine the output of the grammar.

BIOGRAPHICAL SKETCH

Katherine Marie Blake was born in 1995 in Cincinnati, Ohio, and grew up in Carmel, Indiana. She has loved learning languages since starting Spanish in elementary school, and later French in middle school. During a high school immersion program in Saumur, France was when she took her first linguistics course. She graduated with Highest Distinction from Indiana University, Bloomington in December 2016 with Bachelor's Degrees in French and Linguistics (with Honors). She began her PhD in Linguistics at Cornell University in 2017.

To Max.

ACKNOWLEDGEMENTS

First and foremost, I want to thank my advisor, Marten van Schijndel for his support and guidance throughout this project. Marty was not at Cornell when I entered the program in the Fall of 2017, so getting to work with him on my second qualifying paper, and later as the chair of my dissertation, was one of many unexpected joys of graduate school. I greatly appreciated his dry sense of humor, his attentiveness during our meetings, and his thought-provoking alternative hypotheses. His expertise as a computational linguist along with his non-phonologist perspective made this dissertation much stronger, and helped me become a better researcher.

A sincere thank you also goes to my committee members, Abby Cohn, Draga Zec, and Helena Aparicio. Abby is the reason I came to Cornell, and I have greatly enjoyed working with her and learning from her these past five years, from my first year phonology course, to my first qualifying paper, to World Languages Day at Cornell, to my dissertation. She has consistently helped me to see the broader implications of my work and highlight the value of my contributions. I also had the pleasure of working with Draga in several capacities. She is the only faculty member who was on every one of my committees, and I want to thank her for always being a supporter of my work, for asking me interesting questions, and for chatting with me about life and baking. Draga and I also got to teach Introduction to Phonetics and Phonology together for two semesters. I loved teaching this course and getting to share my P-sider excitement with so many great students, and Draga and I made such a great team. The wealth of knowledge that both Abby and Draga have about phonology has been so beneficial for me and for this dissertation. Finally, I want to thank Helena not only for her statistical expertise, but also for her kindness. She showed a sincere interest

in the methodological work of this dissertation, and in me as a person.

Outside of my committee, there are several faculty and staff in the Cornell Linguistics department that helped me along the way. I want to thank Miloje Despic who was the chair of my second qualifying paper, which went on to be the foundation of this dissertation. I learned so much from his syntax expertise, but I also just really enjoyed spending time chatting with him. Bruce McKee also deserves a huge thank you for all he does for the PLab students. He never hesitates to help us in times of need, and he spent many hours solving technical difficulties with me during my first qualifying paper. Thank you also to Sam Tilsen, who served on my first qualifying paper committee, but also helped to create the PLab community which has been my academic and social home for the last five years. Finally, all the hardships and bureaucracy of graduate school would have been much more difficult to navigate without the hard work and kindness of Jenny Tindall and Gretchen Ryan.

A special thank you goes to those who served as language consultants for this dissertation: Francesco Burroni, Italian; Gaja Jarosz, Polish; Sidharth Ranjan and Bhavya Pant, Hindi; and Zahra Alzebaidi and Hassan Munshi, Arabic. I deeply appreciate all the time and expertise they shared with me. This dissertation would not be what it is without their invaluable input.

Now I get to acknowledge the wonderful friends who made graduate school much more bearable. First, I want to thank the grad students at UMass Linguistics who made me feel like a part of their department and made me often wish I could be in two places at once, especially: Maggie Baird, Kaden Holladay, Shay Hucklebridge, Seoyoung Kim, and Andrew Lamont. Next are fellow Cornell grad students who I met through my TAships in French, especially: Peter Caswell, Marie Lambert, and Joe Zappa. I have enjoyed your non-linguist per-

spectives and friendship so much. Finally, my fellow Cornell linguists. Thank you to everyone in the PLab for all of your feedback and support over the years; I am endlessly thankful for the helpful and positive environment we created together. Thank you to Mia Gong, who has been helping me understand syntax since our first year. Thank you to Siree Maspong and Francesco Burroni, who were not only incredible mentors to me, but really dear friends. Thank you to Frances Sobolak for being my roommate and my rock through so many highs and lows. And, thank you to Chloe Kwon and Seung-Eun Kim who showed me true and incredible friendship. Words cannot express how grateful I am for my fellow graduate students whose kindness and support were the greatest unexpected joy of my PhD.

I also want to acknowledge those who contributed to my development as a linguist before I came to Cornell. I am grateful for the Indiana University Honors Program in Foreign Languages (IUHPFL) which exposed me to linguistics for the first time when I was a high school student. I am deeply indebted to the Indiana University Linguistics department, the members of the PhonLab, and especially to: Bob Botne, Stuart Davis, Ken de Jong, and Yoshi Kitagawa, who were the exceptional educators that taught my foundational linguistics courses, and beyond. A special thank you goes to Kelly Berkson, who is the reason I became so interested in phonetics and phonology, the person who introduced me to research and experimental linguistics, and the best mentor I could have asked for.

Thank you so much to my family for supporting me throughout my many, many years of school. I am so grateful that I grew up in an environment where I was encouraged to pursue what I was most interested in. Thank you to my mom for her incredible support and sacrifice; she has been my biggest cheerleader

since day one. Thank you to my dad for his encouragement to do the things I wanted to do, become the person I wanted to be, and with hard work and persistence, to achieve something that has never been done before! Thank you to my uncle, my nana, and my brother who have always been so proud of me and have sought to understood what exactly it is I do. And thank you to Sarah Buchanan, who has been a dear and supportive friend from near and (too) far for the past eight years.

My deepest gratitude is reserved for my fiancé, Max Nelson. Our relationship started as colleagues when we met as undergraduate research assistants in the PhonLab at IU, and only got better from there. It is impossible to thank him enough for how much he's done to support me throughout our six-plus years together, all while completing his own PhD. I truly could not have made it through without him.

This dissertation was principally written in: the Cornell PLab; the UMass Linguistics department; my apartment in Ithaca, NY; Max's apartment in Northampton, MA; Ithaca Bakery, Triphammer in Ithaca; and Haymarket Cafe, Familiars Coffee, and Catalpa Coffee in Northampton. The works of Tchaikovsky provided the main soundtrack, especially Serenade for Strings in C, Op. 48: 1. Pezzo in forma di Sonatina.

TABLE OF CONTENTS

Biographical Sketch . . . iii
Dedication . . . iv
Acknowledgements . . . v
Table of Contents . . . ix
List of Tables . . . xi
List of Figures . . . xv

1 Introduction **1**
1.1 Word order . . . 4
1.1.1 Noun-adjective order . . . 5
1.2 Phonologically-conditioned syntax . . . 11
1.3 Phonological markedness constraints . . . 17
1.4 Methodology . . . 27
1.4.1 Corpora . . . 28
1.4.2 Statistical modeling . . . 31
1.4.3 Semantic clustering . . . 33
1.4.4 Acoustic analysis . . . 35
1.4.5 Summary of hypotheses and predictions . . . 36
1.5 Overview of dissertation . . . 38

2 French **39**
2.1 Noun-adjective flexibility . . . 40
2.2 Phonology . . . 42
2.2.1 Vowel hiatus . . . 43
2.2.2 Voice-disagreeing clusters . . . 46
2.2.3 OCP-Place . . . 47
2.2.4 Length . . . 50
2.2.5 Stress constraints . . . 50
2.3 Methods . . . 51
2.4 Results . . . 55
2.4.1 Regression models . . . 55
2.4.2 Acoustic Sample . . . 59
2.5 Discussion . . . 63
2.5.1 Regression models . . . 63
2.5.2 Acoustic sample . . . 65

3 Italian **66**
3.1 Noun-adjective flexibility . . . 67
3.2 Phonology . . . 69
3.2.1 Stress clash . . . 70
3.2.2 Stress lapse . . . 73
3.2.3 Vowel hiatus . . . 75

3.2.4 Length . . . 77
3.2.5 Consonant cluster constraints . . . 78
3.3 Methods . . . 79
3.4 Results . . . 82
3.4.1 Regression models . . . 82
3.4.2 Acoustic sample . . . 85
3.5 Discussion . . . 89
3.5.1 Regression models . . . 89
3.5.2 Acoustic sample . . . 92

4 Polish, Hindi, and Arabic **94**
4.1 Noun-adjective flexibility . . . 95
4.2 Phonology . . . 97
4.2.1 Stress clash . . . 98
4.2.2 Stress lapse . . . 99
4.2.3 Vowel hiatus . . . 99
4.2.4 Voice-disagreeing clusters . . . 100
4.2.5 OCP-Place . . . 101
4.2.6 Length . . . 102
4.3 Methods . . . 103
4.3.1 Polish . . . 103
4.3.2 Hindi . . . 104
4.3.3 Arabic . . . 106
4.4 Results and discussion . . . 107
4.4.1 Polish . . . 108
4.4.2 Hindi . . . 111
4.4.3 Arabic . . . 113

5 Discussion and Conclusions **116**
5.1 Results summary . . . 116
5.2 Implications for phonology at its interfaces . . . 117
5.2.1 Phonological nature of the effects . . . 117
5.2.2 Impact of additional acoustic and semantic analyses . . . 120
5.2.3 Amendments to the original hypotheses . . . 122
5.2.4 The syntax-phonology interface . . . 125
5.3 Limitations and future directions . . . 130
5.4 Conclusions . . . 134

LIST OF TABLES

1.1 Differences in semantic readings of indirect modification adjectives versus direct modification adjectives, which are ordered: Det > IM > DM > N (Cinque, 2010 p.17). 8
1.2 The Prosodic Hierarchy as proposed by Selkirk (1980) (includes Accentual Phrase, but not Clitic Group) and Nespor and Vogel (1986) (includes Clitic Group, but not Accentual Phrase), and the prosodic level of the constraints examined in this dissertation. 12
1.3 Predictions for the effects of LENGTH on noun-adjective ordering in the languages studied. 17
1.4 Phonological constraints included in the corpus analysis. 19
1.5 Predictions for the effects of CLASH on noun-adjective ordering in the languages studied. 21
1.6 Predictions for the effects of LAPSE on noun-adjective ordering in the languages studied. 22
1.7 Predictions for the effects of HIATUS on noun-adjective ordering in the languages studied. 24
1.8 Predictions for the effects of VOICE on noun-adjective ordering in the languages studied. 25
1.9 Predictions for the effects of OCP-PLACE on noun-adjective ordering in the languages studied. 27
1.10 Common Voice corpus details by language. 29
1.11 Lexical database details by language. 30
1.12 Part-of-speech tagging model details by language. 30
1.13 Example of coding schema for two noun-adjective pairs in Italian. 32

2.1 Adjective types in French. 40
2.2 Token and type frequencies of noun-adjective pairs, adjectives, and nouns in the Common Voice French corpus data. Flexible indicates that the pair or lexical item appeared in both positions, PRENOMINAL and POSTNOMINAL. 41
2.3 Consonant inventory of French (Fougeron and Smith, 1993). . . . 42
2.4 Vowel inventory of French (Fougeron and Smith, 1993). 43
2.5 Proportion of adjective tokens (n=130513) and types (n=7657) that are vowel initial or vowel final in the French Common Voice corpus. 45
2.6 Proportion of noun tokens (n=130513) and types (n=9635) that are vowel initial or vowel final in the French Common Voice corpus. 46
2.7 Violations of HIATUS among noun-adjective pairs (n=130513) in the French Common Voice corpus. 46
2.8 Proportion of adjective tokens (n=130513) and types (n=7657) that have a voiced onset, voiceless onset, voiced coda, or voiceless coda in the French Common Voice corpus. 47

2.9 Proportion of noun tokens (n=130513) and types (n=9635) that have a voiced onset, voiceless onset, voiced coda, or voiceless coda in the French Common Voice corpus. 48
2.10 Violations of VOICE among noun-adjective pairs (n=130513) in the French Common Voice corpus. 48
2.11 Proportion of adjective tokens (n=130513) and types (n=7657) that have an onset or coda at the attested places of articulation in the French Common Voice corpus. 49
2.12 Proportion of noun tokens (n=130513) and types (n=9635) that have an onset or coda at the attested places of articulation in the French Common Voice corpus. 49
2.13 Violations of OCP among noun-adjective pairs (n=130513) in the French Common Voice corpus. 49
2.14 Mean, median, and mode syllable counts for adjectives in the French Common Voice corpus. 51
2.15 Mean, median, and mode syllable counts for nouns in the French Common Voice corpus. 51
2.16 Violations of LENGTH among noun-adjective pairs (n=130513) in the French Common Voice corpus. 51
2.17 Summary table of which phonological constraints are active in French. 52
2.18 Definition of phonological constraints for French. 53
2.19 Cosine similarity values for adjectives with position-specific definitions in *Le Petit Robert*. Those above the threshold of 0.47 are in bold, showing that they are incorrectly categorized. 55
2.20 Model fit for French data containing *less similar* adjectives. Number of observations is 61,060 noun-adjective pairs. $R^2 = 0.41$. . . . 56
2.21 Model fit for French data containing *more similar* adjectives. Number of observations is 34,259 noun-adjective pair tokens. $R^2 = 0.55$. 57
2.22 Results summary for the acoustic analysis of the sample of tolerated voice-disagreeing clusters in French. 62

3.1 Adjective types in Italian. 67
3.2 Token and type frequencies of noun-adjective pairs, adjectives, and nouns in the Common Voice Italian corpus data. Flexible indicates that the pair or lexical item appeared in both orders, PRENOMINAL and POSTNOMINAL. 68
3.3 Consonant inventory of Italian (Kramer, 2009). 69
3.4 Vowel inventory of Italian (Kramer, 2009). 70
3.5 Proportion of adjective tokens (n= 72841) and types (n=8705) that have initial, final, or penultimate stress in the Italian Common Voice corpus. 72

3.6 Proportion of noun tokens (n=72841) and types (n=8715) that have initial, final, or penultimate stress in the Italian Common Voice corpus. . . . 72
3.7 Violations of Clash among noun-adjective pairs (n=72841) in the Italian Common Voice corpus. . . . 72
3.8 Proportion of adjective tokens (n=72841) and types (n=8705) that have initial, final, or penultimate stress in the Italian Common Voice corpus. . . . 74
3.9 Proportion of noun tokens (n=72841) and types (n=8715) that have initial, final, or penultimate stress in the Italian Common Voice corpus. . . . 74
3.10 Violations of Lapse among noun-adjective pairs (n=72841) in the Italian Common Voice corpus. . . . 74
3.11 Proportion of adjective tokens (n=72841) and types (n=8705) that are vowel initial or vowel final in the Italian Common Voice corpus. . . . 76
3.12 Proportion of noun tokens (n=72841) and types (n=8715) that are vowel initial or vowel final in the Italian Common Voice corpus. 76
3.13 Violations of Hiatus among noun-adjective pairs (n=72841) in the Italian Common Voice corpus. . . . 77
3.14 Mean, median, and mode syllable counts for adjectives in the Italian Common Voice corpus. . . . 78
3.15 Mean, median, and mode syllable counts for nouns in the Italian Common Voice corpus. . . . 78
3.16 Violations of Length among noun-adjective pairs (n=72841) in the Italian Common Voice corpus. . . . 78
3.17 Summary table of which phonological constraints are active in Italian. . . . 79
3.18 Definition of phonological constraints for Italian. . . . 80
3.19 Model fit for Italian data containing *less similar* adjectives. Number of observations is 39,568 noun-adjective pairs. . . . 83
3.20 Model fit for Italian data containing *more similar* adjectives. Number of observations is 9,606 noun-adjective pairs. . . . 84
3.21 Results of a one-tailed Z-test for Italian that indicate whether or not the coefficient in the similar model is greater than the same coefficient in the dissimilar model. . . . 85
3.22 Results summary for the acoustic analysis of the sample of tolerated clashes in Italian. . . . 88

4.1 Summary table of which phonological constraints are active in Polish, Hindi, and Arabic. . . . 103
4.2 Definition of phonological constraints for Polish. . . . 104
4.3 Definition of phonological constraints for Hindi. . . . 105
4.4 Definition of phonological constraints for Arabic. . . . 107

4.5 Model fit for Polish data containing flexible adjectives. Number of observations is 9,601 noun-adjective pairs. 109
4.6 Model fit for Hindi data containing flexible adjectives. Number of observations is 242 noun-adjective pairs. Model did not converge. 112
4.7 Model fit for Arabic data containing flexible adjectives. Number of observations is 7,606 noun-adjective pairs. 113

5.1 Summary table of predictions (left) and results (right) in each language for the phonological constraints that are actively avoided (✓) via word-ordering or ignored/tolerated (✗). Parentheses in the table indicate constraints where only one of the simple effects supported the hypothesis. 117
5.2 Example of variation + filtering theoretic grammar (Anttila, 2016). Syntactic constituents are indicated in square brackets and phonological constituents by parentheses. 127

LIST OF FIGURES

1.1 Schematic visualization of the cosine similarity values of flexible adjectives. Two distributions are found in the data, and the boundary between them (dashed line) is used to bin adjectives into *less similar* (to the left) and *more similar* groups (to the right). 34

2.1 Distribution of cosine similarities measured between prenominal and postnominal embeddings of adjectives in French. Cutoff of 0.47 is marked with a red vertical line. 54
2.2 Visualization of the mixed-effects logistic regression results in French of the *dissimilar* (blue, triangle) and *similar* (red, circle) models. Coefficient values with standard error are shown for each fixed effect; significant effects are indicated by a star. 58
2.3 Full assimilation of /z/ to [s] in the phrase *nombreuses fleurs* 'many flowers', spoken by a male French speaker in the Common Voice corpus (common_voice_fr_20269979.mp3). Pitch is tracked in blue and intensity in yellow in the spectrogram. 60
2.4 Partial assimilation of /ʁ/ to [χ] in the phrase *rédacteur principal* 'main editor', spoken by a male French speaker in the Common Voice corpus (common_voice_fr_19812631.mp3). Pitch is tracked in blue and intensity in yellow in the spectrogram. 60
2.5 Pronunciation of an *e muet* in *mystérieuse femme* 'mysterious woman', spoken by a female French speaker in the Common Voice corpus (common_voice_fr_23892094.mp3). Pitch is tracked in blue and intensity in yellow in the spectrogram. 61
2.6 No assimilation of /ʃ/ in the phrase *proche vallée* 'nearby valley', spoken by a male French speaker in the Common Voice corpus (common_voice_fr_19781758.mp3). Pitch is tracked in blue and intensity in yellow in the spectrogram. 62

3.1 Distribution of cosine similarities measured between prenominal and postnominal embeddings of adjectives in Italian. Cutoff of 0.51 is marked with a red vertical line. 81
3.2 Visualization of the mixed-effects logistic regression results in Italian of the *dissimlar* (blue, triangle) and *similar* (red, circle) models. Coefficient values with standard error are shown for each fixed effect; significant effects are indicated by a star. 85
3.3 Stress shift in the noun-adjective phrase *città ricca* 'rich city', spoken by a male Italian speaker in the Common Voice corpus (file common_voice_it_19870780.mp3). Pitch is tracked in blue and intensity in yellow in the spectrogram. 86

3.4 Onset lengthening in the phrase *città sola* 'lonely city', spoken by a male Italian speaker in the Common Voice corpus (file common_voice_it_25962916.mp3). Pitch is tracked in blue and intensity in yellow in the spectrogram. 87
3.5 Final syllable of *miglior*(*e*) produced in the phrase *migliore film* 'best film', spoken by a female Italian speaker in the Common Voice corpus (file common_voice_it_20306010.mp3). Pitch is tracked in blue and intensity in yellow in the spectrogram. 88

4.1 Visualization of the mixed-effects logistic regression results in Polish. Coefficient values with standard error are shown for each fixed effect; significant effects are indicated by a star. 109
4.2 Token frequencies of adjective types in Hindi. 242 total tokens and 33 types. 112
4.3 Visualization of the mixed-effects logistic regression results in Arabic. Coefficient values with standard error are shown for each fixed effect; significant effects are indicated by a star. 114

5.1 Conceptualization of variation + filtering theory, adapted from Anttila (2016). 127
5.2 Conceptualization of the relative strength of phonological effects, which have been theorized to weaken across boundaries and with the presence of non-phonological constraints (Shih, 2014; Martin, 2011). 130

CHAPTER 1

INTRODUCTION

This dissertation investigates the effects of phonological markedness avoidance on noun-adjective ordering. Traditional theories regard syntax as the foundational component of the grammar, generating the abstract form of the utterance whose phonetic form is determined by the phonological component and whose interpretation is determined by the semantic component (Chomsky, 1965). The phonological and semantic components are merely *interpretive*, while the syntactic component is *generative*. The feed-forward structure of this model predicts that the word order determined by the syntactic rules is not affected by the semantic interpretation or phonological representation. Each of these interpretive components accepts the syntactic input and operates on this *deep structure* to produce the *surface structure*.

I contribute to a body of work challenging this idea (e.g., Breiss and Hayes, 2020; Shih, 2014; Zec and Inkelas, 1990) by providing evidence that word ordering is at least partially conditioned by the avoidance of phonologically-marked structures. Using corpora of spoken language, I show that in several languages where noun-adjective ordering is flexible, orders that avoid a phonologically-marked structure at the word boundary between noun and adjective are preferred. Noun-adjective ordering is a key place to look for phonological effects on sentence structure because the possible orders are constrained to either prenominal ADJECTIVE NOUN or postnominal NOUN ADJECTIVE, thus limiting the potential phonological and semantic effects. A noun and its modifying adjective also form a single phonological phrase, eliminating a potential confound at the phrase boundary.

I present case studies of five languages, four of which are Indo-European: Italian and French (Romance), as well as Polish (Slavic) and Hindi (Indo-Aryan); and one Afro-Asiatic language, Modern Standard Arabic (West Semitic). I analyze spoken corpus data in each language to investigate if variation in noun-adjective ordering is conditioned by the avoidance of phonological markedness. Languages were chosen based on flexibility of noun-adjective ordering and availability of high-quality and high-quantity[1] spoken corpora. I examined a set of phonological constraints across all languages, which included constraints targeting prosody, syllable structure, and phonological features. I also report in-depth semantic and acoustic analyses for Italian and French to provide a better context for the semantic factors and the phonetic consequences of noun-adjective ordering.

Using these general methods, this dissertation asks two central questions. First, is variation in noun-adjective ordering conditioned by phonological markedness avoidance? Phonological effects have been found for a variety of syntactic structures in other languages like English (e.g., Breiss and Hayes, 2020; Morgan, 2016), Serbo-Croatian (Inkelas and Zec, 1995), and Sanskrit (Gunkel and Ryan, 2011). Effects on noun-adjective ordering specifically have been found for Tagalog (Shih and Zuraw, 2017), but work on other languages is lacking. I predict phonological markedness avoidance effects on noun-adjective ordering, and provide evidence from five languages.

Second, if there is evidence for this conditioning, does it seem to be a reflection of the phonological grammar that is active elsewhere in the language? I tested a single set of phonological constraints across all languages, but not every constraint is repaired phonologically in each language, meaning that there is not

[1]Except for Hindi, see Chapter 4.

always an active phonological process such as deletion, epenthesis, or assimilation that works to prevent the phenomenon from surfacing. For instance, vowel hiatus is tolerated in Italian to a greater degree than in French which repairs hiatus phonologically with liaison consonants that surface when hiatus arises between words; therefore the hiatus constraint is active in French because it has a phonological repair, whereas this is not the case in Italian. I adopt the hypothesis that only the constraints that are active in a language's phonology will have an effect on word ordering. This follows from results presented by Shih and Zuraw (2017) on Tagalog, where cross-linguistically marked structures, like a sequence of consonants with shared velar place of articulation, that are not penalized in Tagalog were also not found to have an effect on noun-adjective ordering.

Additionally, there has been a lot of discussion of the type of phonological material that is accessible to the syntactic grammar: some have argued that the interface between phonology and syntax is only at the prosodic level (see Inkelas and Zec, 1995), thus segmental effects alone would not be expected under this prosodic hypothesis. The constraint set examined here includes markedness constraints that target prosody, syllable structure, and phonological features to test these effects. I also include a constraint for Fixed Expressionness, which is a numerical approximation of the immovability of a particular noun-adjective phrase in terms of prenominal versus postnominal ordering (Morgan, 2016).

I test the hypotheses of this dissertation with a corpus study. Using large speech corpora, the question of which phonological constraints affect syntactic structure and to what degree is analyzed across several languages. Using mixed effects logistic regression models, comparisons of the predictors in these models show that phonological markedness significantly conditions the surface order-

ing of noun-adjective pairs, following the hypothesis that phonological form has an effect on word order. In French and Italian, the phonetic realization of two of these phonological effects is probed with acoustic analysis, and the semantic effects on word ordering are more deeply examined using vector semantics.

The remainder of this chapter is organized as follows: Section 1.1 is a review of the literature on word order, specifically noun-adjective ordering and gives an outline of the syntactic framework assumed here. Section 1.2 is a review of the literature on the syntax-phonology interface. In Section 1.3, I present the phonologically-marked structures analyzed in this dissertation. Section 1.4 details the general methodology of the studies. Finally, an overview of the remainder of the dissertation is provided in Section 1.5.

1.1 Word order

Languages vary in the extent to which they allow variation in constituent ordering. English is at the relatively strict end of the spectrum, where subject-verb-object is the largely dominant and fixed word order. At the other end of the spectrum is Russian, where the word order comparatively freer. The tendency for languages with rich case-marking systems to have freer word order in comparison to languages with little case marking has been well documented since Sapir (1921).

For example, English has relatively little case marking and strict word order, while Russian has a rich case marking system and enjoys comparatively freer word order; this variation is shown below. Three of the languages studied here: Polish, Hindi, and Arabic have fairly extensive case-marking and free

word order, while French and Italian are more constrained but still exhibit some flexibility.

(1) **English**

a. Jordan likes Casey (SVO)

b. * Casey likes Jordan (OVS)

≠ 'Jordan likes Casey'

c. * likes Casey Jordan (VOS)

(2) **Russian**

a. Masha lubit Petyu (SVO)
Mary.Nom love.3PS Peter.Acc

b. Petyu lubit Masha (OVS)
Peter.Acc love.3PS Mary.Nom

c. Lubit Petyu Masha (VOS)
love.3PS Peter.Acc Mary.Nom

'Mary loves Peter'

While much of the literature on word order flexibility focuses on the major constituents, subject, verb, and object, I analyze a smaller syntactic group, the noun phrase (NP); more specifically, the ordering of noun and adjective.

1.1.1 Noun-adjective order

Typologically, postnominal NOUN ADJECTIVE order is more common than prenominal ADJECTIVE NOUN: of the languages in the WALS database (Dryer and Haspelmath, 2013) with documentation for this ordering, 879 have a dominant postnominal order (64%), 373 have a dominant prenominal order (27%), 110 have no

dominant order (8%), and 5 have only internally-headed relative clauses (1%). The postnominal and prenominal orders seem to be geographically distributed: postnominal order is generally found in Africa, southwestern Europe, and the Middle East, as well as India, Southeast Asia, and the Pacific; prenominal order is generally found in Europe and Asia except for the postnominal areas noted here. The dominant orders for the languages studied here are as follows: French, Italian, and Arabic have largely postnominal NOUN ADJECTIVE ordering and Polish and Hindi have largely prenominal ADJECTIVE NOUN ordering (Laenzlinger, 2005; Hall, 1948; Ryding, 2005; Sadowska, 2012; Jain, 1995). Examples of these two orders are shown in (3). These five languages exhibit some degree of flexibility of this ordering, and this thesis shows that some of this variation can be accounted for by phonological factors. I attempted to include languages with no dominant noun-adjective order, but was constrained by the corpus data available at the time.

(3) **French** and **Polish**

a. la voiture rouge
the car red

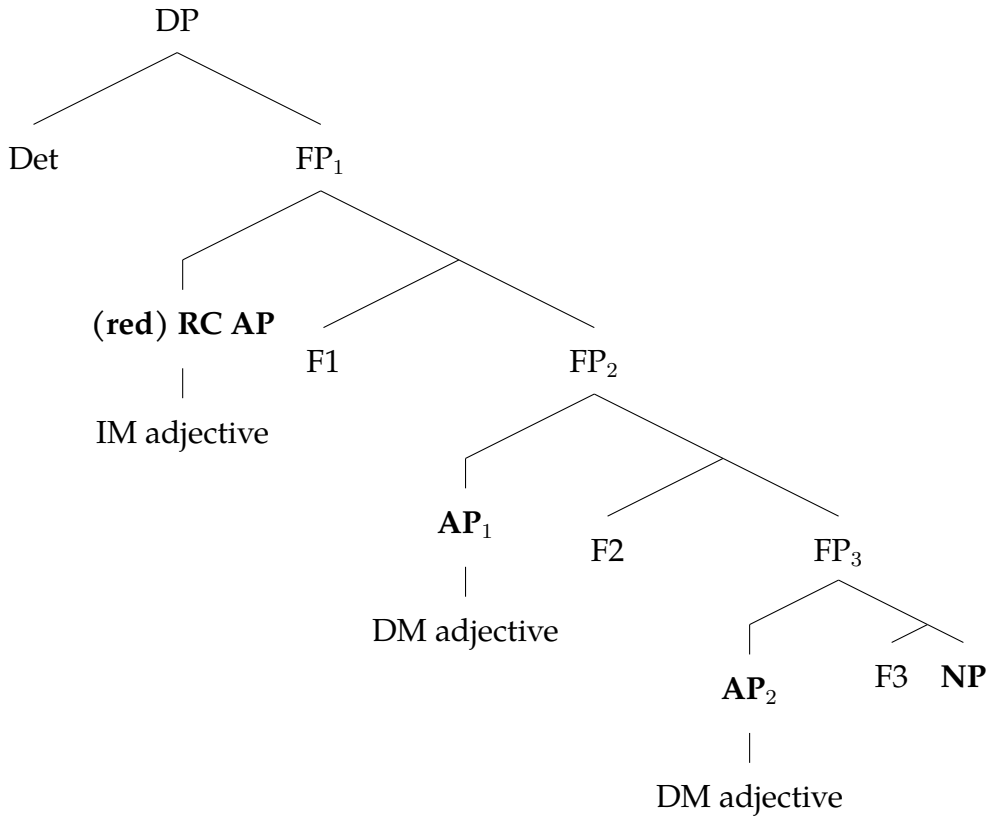
b. czerwony samochód
red car
'the red car'

This dissertation focuses on the language-specific phonological influences that produce different ordering preferences; however, I also begin to address the syntax and semantics of these orderings by including languages with postnominal- and prenominal-dominant orders, and by analyzing noun-adjective pairs with semantically-similar adjectives separately from semantically-dissimilar ones. In regards to the syntax and semantics of adjec-

tives, I assume the approach of Cinque (2010). Under this theory, there are two different types of adjectives, direct and indirect modifiers. This proposal can account for both prenominal and postnominal ordering with XP movement. Both types of adjectives are base-generated prenominally, and different surface orders arise via XP movement of the NP. Direct modification (DM) adjectives come from a functional projection and are non-predicative, and indirect modification (IM) adjectives come from reduced relative clauses and are predicative. IM adjectives are generated higher than DM adjectives in the underlying syntactic structure, and there are semantic differences between the two types, shown in Table 1.1.

Languages with dominant prenominal order, like Polish and Hindi, have no XP raising; languages with dominant postnominal order, like the Romance languages and Arabic, have cyclic XP raising starting with the NP over the first AP, then the XP governing the NP and the now-postnominal AP raising again, and so on. This type of raising accounts for the mirror ordering of multiple adjectives seen in prenominal languages versus postnominal languages (Cinque, 2010).

(4)



Cinque (2010)

	Indirect modification	Direct modification	
Determiner	stage-level	individual-level	Noun
	restrictive	non-restrictive	
	implicit relative clause	modal	
	intersective	non-intersective	
	relative	absolute	
	non-specificity inducing	specificity inducing	
	epistemic 'unknown'	evaluative 'unknown'	
	discourse anaphoric 'different'	NP-dependent 'different'	

Table 1.1: Differences in semantic readings of indirect modification adjectives versus direct modification adjectives, which are ordered: Det > IM > DM > N (Cinque, 2010 p.17).

Importantly, this type of cyclic movement accounts for the semantic ambiguity found in different adjective positions. In prenominal-dominant languages, the prenominal position of the adjective is ambiguous between the indirect and

direct modification readings, while in postnominal-dominant languages, the postnominal position of the adjective is ambiguous. This is shown in English and Italian for the adjectives *beautiful* and *buono* 'good', which can be ambiguous between their intersective and non-intersective readings depending on the position (Cinque, 2010; pp.9-10). In English, both readings are equally accessible in prenominal order as in (5a), but postnominal ordering in (5b) renders the non-intersective reading less accessible. In Italian, both readings are accessible in postnominal ordering as in (6b), but in prenominal order in (5a) the intersective reading is infelicitous.

(5) **English**

a. Olga is a more beautiful dancer than her instructor

= 'Olga is a dancer who is also prettier than her instructor.'

(intersective)

= 'Olga dances more beautifully than her instructor.'

(non-intersective)

b. Olga is a dancer more beautiful than her instructor.

= 'Olga is a dancer who is also prettier than her instructor.'

(intersective)

?= 'Olga dances more beautifully than her instructor.'

(non-intersective) Cinque (2010)

(6) **Italian**

a. Un buon attaccante non farebbe mai una cosa del genere
a good forward not would-do never a thing of-the kind
= 'A forward good at playing forward would never do such a thing.'

(non-intersective)

≠ 'A good-hearted forward would never do such a thing.'

(intersective)

b. Un attaccante buono non farebbe mai una cosa del genere
a forward good not would-do never a thing of-the kind
= 'A forward good at playing forward would never do such a thing.'

(non-intersective)

= 'A good-hearted forward would never do such a thing.'

(intersective) Cinque (2010)

All five languages analyzed in this dissertation have flexibility of the position of adjectives relative to the noun, but often this ordering comes with semantic baggage, similar to the above examples. Languages and even specific adjectives vary in the extent to which the semantic reading of the phrase differs between the two orders. This difference ranges from the phrase constituting a fixed expression in one order, to a shift in emphasis, to no change at all. Accounting for this semantic effect on the variation in word ordering is discussed in detail in section 1.4.3. Alternative syntactic and semantic approaches to adjectives are discussed in section 5.3.

A note on multiple-adjective ordering

Much work has been devoted to describing and theoretically accounting for the ordering of adjectives relative to each other and how this ordering differs between languages (see Cinque, 2010 and references therein). Multiple-adjective ordering is largely attributed to semantic categories. Many works provide evidence for the general tendency of speakers to prefer the hierarchy: QUALITY > SIZE > SHAPE > COLOR > PROVENANCE (e.g., Bloomfield, 1933; Lance, 1968; Vendler,

1968; see also, Hahn et al., 2018; Scontras et al., 2019). This hierarchy is mirrored between languages that have a dominant prenominal versus postnominal placement of the adjective (Sproat and Shih, 1991; Cinque, 2010).

This issue of how multiple adjectives may be ordered with respect to each other, and what governs this ordering and its variation is not the focus of this dissertation. I present an analysis of the variation in ordering of (single) adjectives with respect to the noun that they modify. The interaction of multiple adjectives and the noun that they modify in terms of phonological conditioning is left to future research.

1.2 Phonologically-conditioned syntax

The syntax-phonology interface concerns how the syntactic and phonological components of the grammar interact, both in terms of the direction of influence and the kinds of information that can be shared between the two. Theories range from phonology-free syntax, wherein the influence is unidirectional (Zwicky and Pullum, 1986); to bi-directional influence at the interface (Selkirk, 1978), either totally unconstrained or limited to the prosodic domain. I present empirical evidence in this dissertation that word order is, in part, phonologically conditioned.

Previous work that claims that syntax is not free from phonological effects has proposed various methods of constraining how the components can influence each other. Some have argued that syntactic influence is constrained to the level of the XP, and phonological influence is constrained to prosody (Truckenbrodt, 2007; Selkirk et al., 2011; Ishihara, 2015). I present evidence for phonolog-

ical effects on syntax that extend beyond the prosodic hierarchy, targeting not only prosody and syllable organization, but also segments; see Table 1.2. In the remainder of this section, I review previous work showing phonological effects on syntax, and present the length constraint used in my regression analysis.

The Prosodic Hierarchy	
Level	*Constraints*
Utterance	
Intonational Phrase	
Phonological Phrase	
(Clitic Group)	
(Accentual Phrase)	
Phonological Word	Length
Foot	Clash, Lapse
Syllable	Hiatus
Below the hierarchy: Segmental	Voice, OCP

Table 1.2: The Prosodic Hierarchy as proposed by Selkirk (1980) (includes Accentual Phrase, but not Clitic Group) and Nespor and Vogel (1986) (includes Clitic Group, but not Accentual Phrase), and the prosodic level of the constraints examined in this dissertation.

Previous evidence

Phonological effects on other parts of the grammar have been noted in various languages. Inkelas and Zec (1995) present data from Serbo-Croatian and English as evidence for phonological effects on syntax (as described in Zec and Inkelas, 1990). In Serbo-Croatian, topicalization, a syntactic process, can occur only with constituents that are at least a branching phonological phrase. Phrases which consist of a topicalized constituent that is only one phonological word are judged to be ungrammatical. In English, data in support for phonological influence on syntax come from a well-known phenomenon called Heavy NP Shift. In constructions which have a heavy NP, this "shifted" phrase is minimally two

phonological phrases. If the shifted NP is lighter than this minimum, the sentence is judged to be unacceptable. Examples (7) and (8) are recreated from Zec and Inkelas (1990), pp.373-374.

(7) **Serbo-Croatian**

a. $[[\text{Petar}]_{\omega}\ [\text{Petrović}]_{\omega}]_{\text{NP}}$voleo-je Mariju
Peter Petrovic loved-AUX Mary
'Peter Petrovic loved Mary'

b. * $[[\text{Petar}]_{\omega}]_{\text{NP}}$voleo-je Mariju
Peter loved-AUX Mary
'Peter loved Mary'

(8) **English**

a. Mark showed to John $[[\text{some letters}]_{\phi}\ [\text{from Paris}]_{\phi}]_{\text{NP}}$

b. * Mark showed to John $[[\text{some letters}]_{\phi}]_{\text{NP}}$

In his 2011 work, Martin presents evidence for phonological effects on the lexicon. He looks at morpheme-internal phonotactic constraints in Navajo (required sibilant harmony) and English (ban on geminate consonants). In both languages, these constraints are active within morphemes, but are violated across prosodic word boundaries like compounds. For example, a geminate is not allowed within a root in English, but there do exist compounds like *bookkeeper*. While these are legal, Martin shows them to be statistically underrepresented. He proposes this is due to the absence of monomorphemes that have geminates, and since a learner constructs a grammar based on their linguistic input, they will underpredict the number of compounds with geminates because they generalize the rule for monomorphemes to heteromorphemic words.

Breiss and Hayes (2020) examine phonological markedness effects on sentence formation in English by looking at the avoidance of several phonological markedness constraints (including stress clash and vowel hiatus, as is examined in this dissertation) in sentence bigrams. Using a MaxEnt model to diagnose these avoidances, it was found that syntactic choice (i.e., word order) and lexical choice (i.e., synonym selection) were significant strategies for phonological markedness avoidance. Phonological phenomena found to significantly affect syntactic structure in their study include: stress clash, long consonant clusters, sibilant clash, geminates, vowel hiatus, bad sonority, and nasal-voiceless consonant clusters.

In a corpus study, Shih and Zuraw (2017) investigate noun-adjective ordering in Tagalog. Nouns and adjectives can either appear as noun-linker-adjective, or adjective-linker-noun (the default). The linker can appear as *-ng* or *-na*, and the variation is phonologically conditioned. Various phenomena were tested using corpus data, and the authors found that several phonological effects were active in determining word order of adjective-noun. Given that *-ng* ends in a nasal and nasals are allowed in onset position in Tagalog, it was found that obeying the Obligatory Contour Principle (OCP) for nasals (i.e., avoiding a nasal-nasal sequence) was favored over the default prenominal ordering ADJECTIVE NOUN. The avoidance of a nasal before a voiceless consonant was also found to be a factor in word order, as was the avoidance of vowel-vowel sequences (i.e., *-na* before a word with no onset).

Additionally, there is a considerable amount of literature about the phonological conditioning of binomial expressions, mostly in English (Abraham, 1950; Malkiel, 1959; Bolinger, 1962; Gustafsson, 1974; Cooper and Ross, 1975; Allan,

1987; Golenbock, 2000; Benor and Levy, 2006; Mollin, 2012; Mollin, 2013; Morgan, 2016; Ryan, 2019). Binomials are conjoined phrases of the form "X and Y" that vary in flexibility from frozen, *safe and sound/*sound and safe* to quite free, *television and radio/radio and television* (Morgan, 2016). Experiments have revealed that speakers have preferences for one order over the other, and that these preferences are conditioned by phonology (avoidance of stress clash, Bolinger, 1962; shorter word first, lower F2 vowel first, shorter vowel first, more sonorant onset first, Pinker and Birdsong, 1979), as well as frequency (more frequent first, Fenk-Oczlon, 1989) and semantics (such as male-before-female, Wright et al., 2005).

Length constraint

Termed by Kimball (1973), "Heavy NP shift" has been well-cited as a case where phonological effects on syntax can be seen. This is a phenomenon, originally cited in English but one that has been found in several other languages, wherein a noun phrase is preferred by speakers to come later in the utterance due to its length (i.e., number of words or syllables.) In some cases, the surface position of the NP can even render the utterance ungrammatical due to its length.

(9) Only heavy NP can be final in English

a. Mark showed to John [[some letters]$_{\Phi}$ [from Paris]$_{\Phi}$]$_{\text{NP}}$

b. * Mark showed to John [[some letters]$_{\Phi}$]$_{\text{NP}}$

c. Mark showed [[some letters]$_{\Phi}$]$_{\text{NP}}$to John Zec and Inkelas (1990)

Heavy NP shift is part of a larger body of work investigating "prosodic end-weight" effects, which describe the tendency for heavier constituents to come

later in an utterance, all else being equal (Quirk et al., 1972; Ryan, 2019). Heavier weight may be determined by segmental features, such as vowel length, vowel height (lower), sonority (less sonorous), as well as coda length, and number of syllables.

This same effect is probed in this study: is the longer (i.e., having a greater number of syllables[2]) word in a noun-adjective pair preferred finally, and do postnominal adjectives tend to be longer than prenominal ones? These questions are answered using LENGTH as a feature in the logistic regression analysis, and descriptive statistics about average length of pre- versus postnominal adjectives and nouns in each language.

It has been noted that prosodic end-weight effects are non-existent or even reversed in verb-final (right-branching) languages (i.e., languages with SOV or OSV dominant constituent order; Ryan, 2019). Hindi is the only verb-final language among those investigated in this dissertation, having a dominant SOV word order (McGregor, 1977); this constraint is therefore not predicted to have an effect on noun-adjective word ordering in this language. Other languages are expected to show some effect.

Length effects on word order have been found for binomials (Ryan, 2019), and for nouns in noun-adjective ordering in Tagalog (but not adjectives: Shih and Zuraw, 2017).

[2]Much of the psycholinguistics literature measures length in terms of number of characters, e.g., Siyanova-Chanturia et al. (2017). The choice of the syllable, rather than characters or phonemes, to measure length in this dissertation is a theoretical one: not all phonemes are treated as equally-weight bearing in phonology (e.g., Hyman, 1985; Hayes, 1989; Gordon, 1999). Recent neurocognitive work also supports the importance of the syllable: "the sensitivity to syllable rate [is] arguably the most fundamental property of speech perception and production" (Assaneo and Poeppel, 2018; Coupé et al., 2019).

Language	Prediction: Length
Italian	significant
French	significant
Polish	significant
Hindi	not significant
Arabic	significant

Table 1.3: Predictions for the effects of Length on noun-adjective ordering in the languages studied.

1.3 Phonological markedness constraints

This section presents an overview of the phonological markedness constraints that are both (1) possible at the word boundary and (2) phonologically active in the languages analyzed here. This dissertation poses two main questions: first, is variation conditioned by phonological markedness avoidance; and second, if there is evidence for this conditioning, does it seem to be a reflection of the phonological grammar that is active elsewhere in the language?

To answer the second question, I want to be specific about the definition of "markedness" that I use. The concept was originally introduced by Trubetzkoy (1939) but its definition in phonology has greatly expanded since then, and is generally used to refer to phonological elements that are relatively "unnatural" or "complex," and are those that undergo neutralization or trigger assimilation, for example (Rice, 2007). Three general uses of the term in phonology are categorized by Hume (2011): it may refer to "universal" markedness, meaning that it is a cross-linguistic principle that has a hand in language acquisition, change, and processes; "descriptive" markedness, as a tool to highlight asymmetries in sets of linguistic observations; and, in markedness "constraints," under Optimality Theory (OT; Prince and Smolensky, 2004).

In this dissertation, I define phonologically-marked elements as those that are actively avoided by a particular language's phonological grammar. Active avoidance consists of phonological repair strategies of a sequence, such as liaison for vowel hiatus in French and regressive assimilation for voiceless-voiced or voiced-voiceless obstruent clusters in Polish. I consider hiatus to be phonologically-marked in French, as are voice-disagreeing clusters in Polish because the language-specific phonology has a repair process that acts to prevent the sequence from surfacing. In this dissertation, these marked structures and their repairs are taken to operate in an OT-style grammar, and therefore the structures may trigger multiple repair strategies that work to prevent them from surfacing (i.e., conspiracies; see Chapter 5 for a more in-depth discussion of the theoretical implications of these findings).

I predict that only phonologically-marked structures, as I have defined markedness here, will have an effect on noun-adjective word ordering for each language; a phenomenon not otherwise active in a language's phonology will not have ripple effects in that language's syntax. This distinction is important because "markedness" has been defined in the literature at the level of universal grammar, acting as a force that universally guides inventories, synchronic and diachronic processes, acquisition, phonotactics, and relative frequency (Trubetzkoy, 1939; Rice, 2007). Sequences that are "universally" marked as defined by Trubetzkoy, but are not language-specifically active in the phonological grammar as I define markedness, are *not* predicted to have any effects on noun-adjective ordering for each language. Word ordering and phonological repairs may coexist as strategies for avoiding the surfacing of phonological markedness. If word ordering is utilized instead of a phonological repair, then faithfulness in phonological form is maintained. If word order is preserved for structural,

semantic, or stylistic reasons, then a phonological repair may act to prohibit markedness from surfacing.

The following subsections provide background on the phonologically-marked phenomena that I investigate, and outline the predictions of their effect on noun-adjective ordering in each of the languages of study. The phonological descriptions that I present in this section are from previous literature on these languages as they are spoken by the majority of the educated, mainland population (i.e., French as spoken in France but not Canada). This is particularly an issue for Arabic: I present descriptions of Modern Standard Arabic, or regional dialects where noted; the vast variation between dialects, and taking Modern Standard Arabic as the object of study are issues further discussed in Chapter 4.

A summary of the constraints and predictions is provided in the table below. More details about the extent to which these phenomena are found in each language studied are given in their respective chapters.

Constraint	Phonological material	Languages where marked
Stress clash	Prosodic	Italian, (Hindi)
Stress lapse	Prosodic	Italian
Vowel hiatus	Syllabic	French, Polish, Hindi
Voice-disagreeing clusters	Segmental	French, Polish, Hindi, Arabic
OCP-Place	Segmental	Hindi, Arabic

Table 1.4: Phonological constraints included in the corpus analysis.

Stress clash

Stress clash occurs when two prominent syllables are next to each other. The study of stress clash has origins in metrical phonology, where noted stress shifts

in English words were explained by the drive to avoid adjacent stressed syllables, i.e., the "rhythm rule" (Kiparsky, 1966; Liberman, 1975; Liberman and Prince, 1977; Hayes, 1984). The classic example of this comes from the following pair.

(10) a. *thirtéen*

Stress-final when produced in isolation/phrase-finally

b. *thírteen mén*

Stress is shifted leftward before a stressed syllable

Stress shift due to clash has been noted in other languages: German (Kiparsky, 1966), Dutch (Rischel, 1972), Hebrew (Prince, 1975; McCarthy, 2018), Passamaquoddy (Stowell, 1979), Dari (Bing, 1980), Finnish (Hayes, 1980), and importantly for my analysis: Italian (Nespor and Vogel, 1979). Stress shift in Italian due to clash has been widely discussed, and is expected to have an effect on noun-adjective ordering. More details about the rhythm rule in Italian are in Chapter 2.

Stress in French is generally considered to be assigned only at the phrasal level and not at the lexical level, therefore clash as a factor in noun-adjective ordering for this language is not relevant. Word-final stress is not possible in Polish, so clash cannot occur between words[3]. In Hindi, clash between words is possible (i.e., word-initial stress and word-final stress are both found), but its avoidance is reported to be variable (Pandey, 2021). Final stress is possible in Arabic, but very rare and there are no previous reports of clash avoidance. A summary of the predictions for the effect of clash on noun-adjective word ordering in these languages is in Table 1.5.

[3]Except, of course, between monosyllabic words; however, in this case, clash would occur in either ordering of the two words and therefore would not be a conditioning factor on ordering.

Language	Prediction: Clash
Italian	significant
French	not possible
Polish	not possible
Hindi	potentially significant
Arabic	not significant

Table 1.5: Predictions for the effects of Clash on noun-adjective ordering in the languages studied.

Much of the effects of clash on syntactic order and lexical selection come from English. Clash has been found to affect binomial ordering (McDonald et al., 1993; Wright et al., 2005; Benor and Levy, 2006). Schlüter and Knappe (2018) find evidence for stress clash avoidance between adjectives and their following nouns in English via synonym selection of an alternative adjective that does not have final stress before an initially-stressed noun. Clash was found to affect the English dative construction to a small extent by Shih (2017). Speyer (2008) attributes the decline of topicalization and V2 word order in English, in part to clash avoidance. Additional clash effects on syntactic structure in English can be found in Schlüter (2005), pp.35-42.

Stress lapse

Stress lapse occurs when three or more non-stressed syllables are adjacent. Like clash, lapse also works towards an alternating stress pattern of strong and weak syllables, which is broadly preferred by languages (Hayes, 1980; Prince, 1983; Hayes, 1984; Selkirk, 1984; Nespor and Vogel, 1989; Rubach and Booij, 1985; Green and Kenstowicz, 1995).

Lapse is generally repaired by the addition of prominence on a weak syllable

(termed "Beat Addition," Nespor and Vogel, 1989). This has been shown for Greek and Italian, as well as English, which also shows further reduction of the weak string of syllables to minimize the gap between stresses as alternative strategy for lapse repair (Nespor and Vogel, 1989). An example is shown below for Greek, where bracketing indicates clitic groups (Nespor and Vogel, 1989, p.95).

(11) **Greek**

a. [o ðáskalos mu] [to ⌊pe] → o ðáskalós mu to ⌊pe

'My teacher said it'

Predictions for LAPSE parallel those for CLASH: I predict Italian will also show an effect of LAPSE; French does not have word-level stress, so no effect is predicted; and Hindi is not expected to show an effect. While Polish and Arabic do not have final stress, initial stress is found in both, making LAPSE possible. Dispreference of lapse in Polish is unclear, and if so, likely occurs at the level of the foot instead of the syllable (Newlin-Łukowicz, 2012). Because of this, it is not included in the model for Polish. In Palestinian Arabic, three unstressed syllables in word final position trigger a stress shift, an effect of lapse (Houghton, 2008); no such shift has been reported for Modern Standard Arabic, so no effect on word ordering is predicted.

LANGUAGE	PREDICTION: LAPSE
Italian	significant
French	not possible
Polish	not possible
Hindi	not significant
Arabic	not significant

Table 1.6: Predictions for the effects of LAPSE on noun-adjective ordering in the languages studied.

Previous work on the interaction between lapse avoidance and word ordering found an effect on binomials in English (McDonald et al., 1993; Wright et al., 2005; Benor and Levy, 2006; Mollin, 2012).

Vowel hiatus

Vowel hiatus is a sequence of two adjacent vowels, belonging to different syllables. This sequence has generally been considered "unstable" or "marked" in phonology (Kenstowicz, 1994, p.23; Trask, 1996). Cross-linguistic avoidance of hiatus is well-documented, showing many different repair strategies (Casali, 2021). A sequence of two vowels may be repaired with deletion of one of the two vowels, as in Nambya (Kadenge, 2013); with the insertion of an epenthetic consonant, as in Mongolian (Svantesson et al., 2005); with glide formation, as in Kavalan (Lin, 2018); with coalescence of the two vowels, as in Anufo (Adjekum et al., 1993); or with diphthongization, as in Haitian Creole (Fournier, 1978).

Hiatus repair strategies are marginal in Italian, thus I expect no effect on word ordering. Liaison is a well-studied vowel hiatus repair in French; under certain syntactic and register conditions, liaison consonants surface between two words with vowels at the shared boundary (Tranel, 1995). Glottal stop epenthesis repairs hiatus in Polish (Schwartz, 2013). Glide insertion breaks up vowel-vowel sequences in Hindi (Singh and Sarma, 2011; Kachru, 2006; Ohala, 1983). Finally, hiatus at the word boundary is not possible in Arabic, as words are not vowel-initial (Ryding, 2005).

Hiatus has been found to influence word ordering in several languages. Avoidance of hiatus via ordering of noun-adjective pairs was found by Shih and

Language	Prediction: Hiatus
Italian	not significant
French	significant
Polish	significant
Hindi	significant
Arabic	not possible

Table 1.7: Predictions for the effects of Hiatus on noun-adjective ordering in the languages studied.

Zuraw (2017) in a corpus of Tagalog. Gunkel and Ryan (2011) found hiatus avoidance via word ordering of flexible bigrams in the Rigveda, an ancient text written in Vedic Sanskrit. Hiatus is also among the many factors influencing binomial ordering in English (McDonald et al., 1993; Wright et al., 2005; Benor and Levy, 2006; Mollin, 2012).

Voice-disagreeing clusters

This phonotactic constraint restricts consonant clusters whose members do not have the same laryngeal specification, i.e., one consonant is voiced while the other is voiceless. This mismatch is often dispreferred, and may be resolved by (usually regressive) assimilation of voicing of one consonant to the other (Lombardi, 1999)[4].

Word-final consonants are phonotactically disallowed in Italian, and the language seems to lack voicing assimilation in loanwords (Huszthy, 2016). Regressive voicing assimilation has been reported for obstruent clusters in French (Snoeren and Segui, 2003; Hallé and Adda-Decker, 2007) and Polish (Guss-

[4]In Lombardi (1999)'s work, she attributes the cross-linguistically common phenomenon of word-final devoicing to coda neutralization, thus a change in voicing specification is due to syllable wellformedness. While coda neutralization is present in Polish, this issue is orthogonal to that of consonant sequences across the word boundary examined in this dissertation.

mann, 1992), both within and across words. Ohala (1983) reports that consonant clusters that disagree in voicing are disallowed in Hindi. Finally, voicing assimilation is reported in several dialects of Arabic (Egyptian, Sudanese, and Daragözü, Abu-Mansour, 1996; Palestinian, Tamim, 2017; Cairene Kabrah et al., 2011). Given these facts, the predictions about VOICE for the languages of study are in Table 1.8.

LANGUAGE	PREDICTION: VOICE
Italian	not possible
French	significant
Polish	significant
Hindi	significant
Arabic	significant

Table 1.8: Predictions for the effects of VOICE on noun-adjective ordering in the languages studied.

Not much prior work has been done on the avoidance of voice-disagreeing clusters via word ordering. Shih and Zuraw (2017) report avoidance of nasal-voiceless consonant contact in noun-adjective ordering in Tagalog; and, Breiss and Hayes (2020) report some effect of nasal-voiceless consonant contact on sentence formation in English.

OCP-Place

This phonotactic constraint is a restriction on consonant clusters with the same place of articulation, which is considered a violation of the Obligatory Contour Principle. The Obligatory Contour Principle (OCP) was originally proposed within the autosegmental framework of phonology, banning sequences of identical tones in the underlying form (Leben, 1973); and then was extended to a ban on adjacent identical feature specifications (Goldsmith, 1976; McCarthy, 1986).

In addition to OCP-Place effects in cluster reduction in Catalan and Korean, among others (Morales, 1995; Kang, 1998), bans on consonants with adjacent place features have been widely studied in Semitic languages, including Arabic (Greenberg, 1950; Pierrehumbert, 1993; McCarthy, 1994; Frisch and Zawaydeh, 2001). Consonants in the root that are not linked to the same underlying representation cannot share place features; this bans forms like **sasam* but not *samam* 'to poison', as in the latter case the *m*s are argued to be ultimately linked to the same node (Frisch and Zawaydeh, 2001). McCarthy (1986) also describes how metathesis is blocked when it would allow violations of the OCP to surface. Given these facts, an effect of OCP-Place is expected to be found for Arabic at the word boundary between nouns and adjectives.

OCP-Place is irrelevant in this specific structure in Italian, just as is Voice; since the language does not allow word-final consonants, no effects are predicted across the word boundary (Huszthy, 2016). Consonant-consonant sequences across the word boundary occur in French but a dispreference for a shared place of articulation in continental French is unclear (see Côté, 1997 for some claims about Québec French). No previous work on Polish suggests this dispreference either. Finally, an effect is expected in Hindi, where "initially, medially, and finally, two stops of the same point of articulation do not follow each other" (Ohala, 1983, p.56). As far as I know, there is no previous work showing an effect of OCP-Place on word ordering.

Language	Prediction: OCP-Place
Italian	not possible
French	not significant
Polish	not significant
Hindi	significant
Arabic	significant

Table 1.9: Predictions for the effects of OCP-Place on noun-adjective ordering in the languages studied.

1.4 Methodology

This dissertation is primarily a corpus study. Data come from over 1200 hours of speech across the five languages. Examining phonological effects on syntactic structure using this type of data has several advantages. First, the effects examined here are predicted to be subtle in terms of effect size, following previous work (e.g., Morgan, 2016), but are expected to be borne out over the large-scale of data. Semantic and lexical effects are often also at play and may overpower phonological well-formedness to varying degrees. Over a large set of data, these effects are expected to emerge whereas they may be too small to be observed in narrower datasets. Second, because of the vast range of the data, all constraints examined here could be tested at the same time or others could be easily added in follow-up work. Finally, I generalized and released the analysis method I developed here so that it is available for application to additional languages and constraints, making the reproducibility and extension of my work more accessible. All of the scripts developed for this dissertation, which go from corpus data to a dataset with columns for the regression analysis as well as the semantic clustering analysis and more, are publicly available on my GitHub: github.com/katherineblake/dissertation.

The methods I employ are largely based on previous work on binomials (Morgan, 2016; Benor and Levy, 2006) and sentence structure at the bigram (Breiss and Hayes, 2020) and noun-adjective (Shih and Zuraw, 2017) levels. There are a lot of parallels between noun-adjective ordering and binomials ("X and Y"), which along with other syntactic alternations is typically modeled using logistic regression, the main basis of statistical analysis also used here (Morgan, 2016; Benor and Levy, 2006; Bresnan et al., 2007). Similar constraints and descriptive statistics are reported in Breiss and Hayes (2020) and Shih and Zuraw (2017).

1.4.1 Corpora

This thesis analyzes spoken corpus data from the Common Voice corpus, provided by Mozilla[5]. These corpora consist of user-submitted voice recordings; spoken corpora are essential for phonological analysis, so that the acoustic production of an uncontrolled set of data may be analyzed (Cohn and Renwick, 2021). Common Voice is a freely-available and community-built corpus originally designed for building voice recognition systems. In its entirety, it contains over 75 languages and 13,000 hours of recorded speech. Recordings are user-submitted and must be confirmed for intelligibility and accuracy to the orthographic transcription by at least two listeners in order to be validated. Only validated speech was used in this analysis. Details about the versions used for each language are in Table 1.10. Further details about the corpus data for each language are provided in subsequent chapters.

I carried out the majority of the analysis on the phonemic representations of

[5]Accessed Fall 2021 at the following address: voice.mozilla.org

Language	Corpus details
Italian	version it_317h_2021-07-21 288 hours of validated speech 6,407 speakers
French	version fr_834h_2021-07-21 747 hours of validated speech 15,391 speakers
Polish	version pl_152h_2021 129 hours of validated speech 2,918 speakers
Hindi	version hi_11h_2021-07-21 8 hours of validated speech 214 speakers
Arabic	version ar_137h_2021-07-21 85 hours of validated speech 1,052 speakers

Table 1.10: Common Voice corpus details by language.

the audio in Common Voice. Common Voice is transcribed orthographically, so I first converted sentences to the phonemic level using lexical databases. Sentences have text and audio file path pairs, but the audio is not forced aligned. For these reasons, I used lexical databases containing word-level phonemic transcriptions to add pronunciation information to the dataset. This involved looking up the phonemic representation of each orthographic word in the provided database. The phonological information in these databases ranges from automatic grapheme-to-phoneme generation as in WikiPron (Lee et al., 2020), to expert, manual inspection of the entire dataset as in PhonItalia (Goslin et al., 2014). Thus, what is taken as the underlying form of noun-adjective pairs in this analysis is by no means perfect, but is the best approximation available on this scale. Details about these databases are in Table 1.11. Further details about the phonological data for each language are provided in subsequent chapters.

[6]Written by myself with Hassan Munshi, available at github.com/katherineblake/language-scripts. This script generates the phonemic IPA transcription of any Modern Standard Arabic

Language	Lexical Database
Italian	PhonItalia (Goslin et al., 2014) 120,000 words
French	Lexique 3 (New et al., 2004) 140,000 words
Polish	WikiPron (Lee et al., 2020) 86,000 words
Hindi	WikiPron (Lee et al., 2020) 13,000 words
Arabic	Buckwalter-to-IPA script[6]

Table 1.11: Lexical database details by language.

I identified noun-adjective pairs using an automatic part-of-speech tagger. `spaCy` models[7] were used for Italian, French, and Polish. Hindi and Arabic are not supported by `spaCy`, so `Stanza` was used instead for Hindi (Qi et al., 2020)[8], and MADAMIRA for Arabic (Pasha et al., 2014). Exact models and accuracy of their part-of-speech taggers are reported in Table 1.12.

Language	Model
Italian	it_core_news_sm 97% accuracy
French	fr_core_news_sm 96% accuracy
Polish	pl_core_news_sm 98% accuracy
Hindi	Version 1.2.3, Hindi models 98% accuracy
Arabic	MADAMIRA 96% accuracy

Table 1.12: Part-of-speech tagging model details by language.

In addition, I conducted in-depth semantic and acoustic analyses on Italian and French due to the more extensive resources available for these languages.

word transliterated to Buckwalter.

[7] Accessed Fall 2021 at spacy.io/models.

[8] Accessed Fall 2021 at github.com/stanfordnlp/stanza.

The largest corpora and lexical databases are available for these two languages, often by a generous margin. I include Polish, Hindi, and Arabic despite the smaller amount of data available for these languages because of the added linguistic diversity that allows my analysis to generalize outside of the Romance sub-family.

1.4.2 Statistical modeling

Following (Morgan, 2016), I fit a mixed effects logistic regression model using the `glmer` function in R (R Core Team, 2016), predicting the surface word ordering of noun-adjective pairs with a flexible adjective in each language with the constraints outlined in the previous sections: LENGTH, CLASH, LAPSE, HIATUS, VOICE, and OCP-PLACE. An adjective is considered to be flexible if it occurs at least once in both prenominal and postnominal position in the corpus[9]. An additional feature approximating the degree to which the noun-adjective pair is a fixed expression was also used in the model, and is detailed below. The regression analyses are carried out on datasets of noun-adjective pair tokens (non-unique instances of pairs in the corpus) rather than pair types (unique instances of pairs), though descriptive statistics of both are provided in subsequent chapters.

Noun-adjective pair tokens were coded for their ordering outcome: 1 if in prenominal ADJECTIVE NOUN order, 0 if in postnominal NOUN ADJECTIVE order. Phonological constraints were coded in the following way: each pair was coded according to which order, if any, produced a pair that was well-formed for that

[9]It is possible that some adjectives considered flexible by speakers are missed using this method due to their absence in both positions in these particular corpora.

constraint. Constraints are coded as 1 if the pair is well-formed only in prenominal order, as -1 if the pair is well-formed only in postnominal order, or as 0 if well-formedness has no order preference (i.e., both orders are ill- or well-formed.) This coding scheme follows previous work on binomials, Morgan (2016). An example is provided for two pairs in Italian, which has a default postnominal NOUN ADJECTIVE order. All phonological constraints were treated as three-level factors in R, with 0 set as the reference level (i.e., constraints were dummy coded).

DATA	DEPENDENT VARIABLE	INDEPENDENT VARIABLES					
PAIR	ORDER	CLASH	LAPSE	HIATUS	VOICE	OCP-PLACE	LENGTH
/ˈpik.ko.lo/ /ˈal.be.ro/ 'small tree'	1	0	0	-1	0	0	0
/t͡ʃit.ˈta/ /ˈpik.ko.la/ 'small city'	0	1	-1	0	0	0	-1

Table 1.13: Example of coding schema for two noun-adjective pairs in Italian.

Fixed Expressionness constraint

I also included a constraint for relative frequency, calculated for each unordered pair type. This continuous variable is the degree of flexibility of the pair, ranging from 0.5, meaning it occurs in both orders in the dataset evenly, to 1.0, meaning it occurs only in one order. Though it does not account for the semantic cohesiveness of a pair, this constraint is intended to serve as a rough proxy of the degree to which a noun-adjective pair is a fixed expression (Morgan, 2016).

$$\text{Relative frequency} = \left|\frac{\text{prenominal token frequency}}{\text{total token frequency}} - 0.5\right| + 0.5$$

There are many fixed expressions involving adjectives that appear in prenominal and postnominal positions, so degree of lexicalization of any given phrase must also be taken into consideration and balanced against so-called syntactic effects. For example, *pubblico* 'public' occurs in both positions in the Italian corpus data, but must occur postnominally with *sanità* for the phrase 'public health.' Noun-adjective pairs with an adjective that occurred in only one position were excluded from the data, and the fixed-expression constraint was used as a fixed effect in the models to control for the effect of lexicalization of any given noun-adjective pair.

1.4.3 Semantic clustering

The following semantic clustering method was used to control for meaning differences between adjectives in prenominal ADJECTIVE NOUN position and adjectives in postnominal NOUN ADJECTIVE position, as described by Cinque (2010) (see section 1.1 for more details.) Word embeddings were created using the bag-of-words method, meaning that co-occurrences of adjectives and each word in the lexicon were tallied at the sentence level. This was done separately for adjectives occurring in prenominal position and adjectives in postnominal position to capture distributional differences between the two, yielding two adjective-word co-occurrence count matrices. Adjectives that occurred less than two times were removed from each matrix. The filtered count matrices were converted to positive pointwise mutual information (PPMI), and dimensionality was reduced to 128 dimensions using singular value decomposition (SVD). This method follows a relatively standard procedure for computing word embeddings from count information (Schütze, 1993; Bullinaria and Levy, 2007) and has been shown to be

related to the popular modern method, `word2vec` (Mikolov et al., 2013a,b; Levy and Goldberg, 2014; Levy et al., 2015).

$$\text{PPMI}(adj, lexeme) = \max\left(\log \frac{p(adj, lexeme)}{p(adj)p(lexeme)}, 0\right)$$

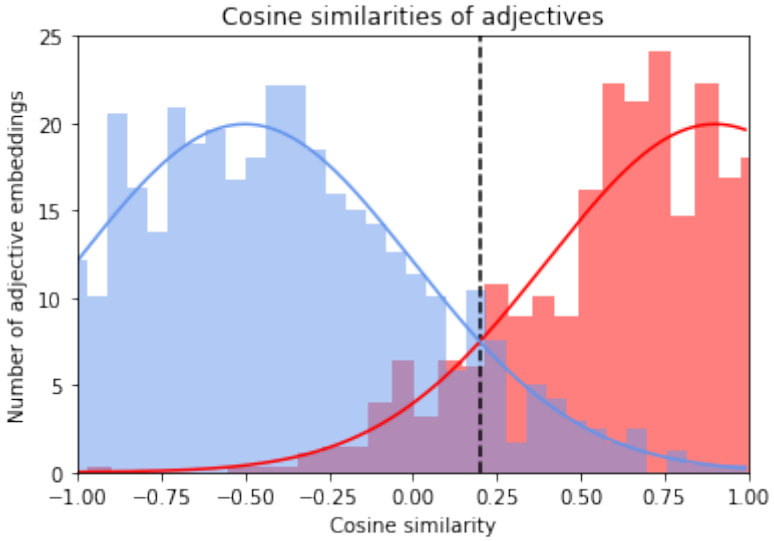


Figure 1.1: Schematic visualization of the cosine similarity values of flexible adjectives. Two distributions are found in the data, and the boundary between them (dashed line) is used to bin adjectives into *less similar* (to the left) and *more similar* groups (to the right).

Cosine similarity was calculated row-wise between the prenominal and postnominal embedding matrices (e.g., between the embedding of *grand* 'big' in prenominal position and the embedding of *grand* 'big' in postnominal position; Salton and Buckley, 1988). A Gaussian mixture model (GMM) with k=2 was fit to the cosine similarities to divide the data into two distributions. The boundary between these two distributions was used to bin the data for the regression models; further details about the determination of this boundary are provided in the methods sections of the relevant chapters (2.3 and 3.3). All noun-

adjective pairs containing an adjective below the boundary were grouped into a *less similar* dataset; and pairs with an adjective above the boundary were grouped into a *more similar* dataset; pairs containing an adjective without a cosine similarity value (i.e., those with an adjective that occurred in only one position) were not included in the subsequent regression analyses. A schematic example of this process is provided in Figure 1.1. Language-specific details about the semantically-grouped datasets are provided in the relevant chapters (2.3 and 3.3). Separate regressions were run on the *less similar* and *more similar* datasets for French and Italian.

In order to test if coefficients in the more similar model are significantly greater than those in the less similar model, a one-tailed Z-test was run on pairs of coefficients that are significant in both models (Clogg et al., 1995). Z-score was then converted to *p*-value for interpretability.

$$\mathrm{Z} = \frac{\beta_1 - \beta_2}{(SE\beta_1)^2 + (SE\beta_2)^2}$$

1.4.4 Acoustic analysis

This dissertation principally investigates the avoidance of marked sequences based on what are taken to be the phonemic forms of lexical items. The question remains, however, what the surface realizations may be of phonologically-marked sequences that are not avoided via word ordering. For example, in Italian, are underlying clashes produced, or are they repaired with stress shift?

In order to address this question, two samples of 50 tokens each were ana-

lyzed in Praat for stress clash in Italian and voice-disagreeing clusters in French. Tokens are comprised of instances where the constraint is tolerated, rather than avoided, at the phonemic level between a flexible noun-adjective pair in the order in which it appears in the corpus. The sets of 50 were chosen based on the recording quality, perceived nativeness of the speaker, and a bias for a diversity of pair types. Tokens were evaluated for presence or absence of the acoustic reality of a phonological repair based on perception and spectrogram information such as intensity, pitch, and duration (clash), and voicing (voice-disagreeing clusters). For example, in the case of an underlying voiceless-voiced consonant sequence across the word boundary, I looked for evidence for assimilation, a phonological repair, present in the acoustic signal via the presence of a voicing bar in the spectrogram during the underlyingly voiceless consonant.

1.4.5 Summary of hypotheses and predictions

The principal hypothesis of this dissertation is that, in addition to semantic and lexical factors, word ordering is phonologically conditioned; therefore, noun-adjective ordering can be influenced by phonological structures. This has been found to some extent in Tagalog by Shih and Zuraw (2017). In this dissertation, I contribute results supporting this hypothesis from five languages.

Given positive evidence for the first hypothesis, I further hypothesize that it is not the case that *all* phonological processes can have this effect. As pointed out in Shih and Zuraw (2017), only the phonological processes that are *active* in a language are predicted to have some influence over the syntax. A phenomenon is active if there is a phonological repair strategy that works against

it, on a language-specific basis. In addition, phonological effects on ordering are predicted to be more prominent as semantic differences between orders are less salient.

Baseline Hypothesis: Noun-adjective ordering can be influenced by phonological markedness.

Phonologically-conditioned Hypothesis: Only those phonologically-marked phenomena that are avoided with phonological repairs may also be avoided with syntactic repairs.

Phonological Prediction: For every phenomenon that is actively phonologically avoided in a language, it may also be avoided syntactically via word-order manipulation in noun-adjective pairs.

Semantically-conditioned Hypothesis: Phonological effects on ordering are stronger if semantic differences between orders are minimal.

Semantic Prediction: Phonological markedness will have a comparatively larger effect on word-ordering in noun-adjective pairs that have less semantic difference between order than on pairs that have a greater difference between orders.

Null Hypothesis: Syntactic variation is independent from phonological influence.

I state the null hypothesis below, which predicts no effects of phonological markedness constraints on flexible noun-adjective ordering. Where word ordering is conditioned by phonology, previous work proposes that the syntactic component produces multiple linearizations which are then filtered by phonological wellformedness (Anttila, 2016). This theory fits well with the results found in

this dissertation (see 5.2.4 for a more in-depth discussion). Using noun-adjective ordering to test the differences in these hypotheses is left to future work. Alternative hypotheses include a grammar where phonology strictly feeds syntax, and a grammar where phonological markedness is repaired with word ordering instead of phonology; both of which would predict no phonological repairs after the ordering has been set. The syntactic and phonological components could also be independent of one another, each assigning probabilities to various outputs, one of which is then produced based on the combined likelihood. Results presented in this dissertation show that phonological repairs take place in both the default and non-default orders, where markedness could have been avoided via word ordering but was not.

1.5 Overview of dissertation

The rest of this dissertation is organized as follows: in Chapter 2, I present the results of the corpus study of French, including additional analyses of semantic effects and phonetic outcomes of noun-adjective word ordering. Voice and Hiatus were found to be avoided. In Chapter 3, I present the results of the corpus study of Italian, including semantic and phonetic analyses. Length was found to be significantly avoided in both models. In Chapter 4, I present the results of the three additional languages, Polish, Hindi, and Arabic. Voice and Length were significantly avoided in Polish and Arabic, in addition to OCP-Place in Arabic. Results were inconclusive for Hindi. In Chapter 5, I summarize the findings of the corpus studies and discuss their impact on our understanding of the syntax-phonology interface.

CHAPTER 2

FRENCH

French is a Romance language (Indo-European), spoken primarily in France and the European Union, with an estimated 274 million speakers worldwide (Local, 2014). This chapter mainly discusses aspects of continental French, a general term for the variety of French widely spoken in France by the educated population. Québécois French is another variety with a substantial population of speakers, largely in Canada. The data from the Common Voice French corpus used here has limited dialect information, and contains speech from over 15,000 speakers. 63% of the speech comes from mainland French speakers, and 5% from speakers in Canada, Belgium, and Switzerland combined. The dialect information of the remaining 32% of the data is not reported[1].

The canonical constituent order of French is Subject-Verb-Object (SVO), but other orders are acceptable (Lahousse and Lamiroy, 2012). Especially relevant to this work is the ordering of adjective and noun, which is canonically NOUN ADJECTIVE (postnominal), but prenominal order is allowed for some adjectives and required for others. This is further elaborated on in the next section. Section 2.3 provides details on methodology specific to the French analysis. Results are presented in section 2.4, and discussed in section 2.5.

[1]With the addition of dialect information for all speakers and enough data, a dialect-specific analysis may further elucidate the results presented in this chapter, and may reveal systematic differences between dialects (Cohn and Renwick, 2021).

2.1 Noun-adjective flexibility

The canonical ordering of adjectives with respect to the noun that they modify is postnominal: NOUN ADJECTIVE. There exist adjectives that can occur only in prenominal position, and those that can occur both before and after the noun. Prenominal position has been described as pertaining to "elementary" adjectives like *petit* 'small', broad or vague adjectives, and quantificational adjectives (Nølke, 1996; Laenzlinger, 2005). Adjectives that can appear in both positions may have a difference in meaning, as described by Cinque (2010) (see Table 1.1), Nølke (1996), and Laenzlinger (2005). Thus, adjectives in French can be thought of as belonging to one of three groups: strictly postnominal, strictly prenominal, or able to appear in both positions. Examples of all three are shown in Table 2.1 (data from Knittel, 2005).

TYPE	FRENCH
(1) Strictly prenominal	??*une maison* ***belle*** $\sim$ *une* ***belle*** *maison* 'a nice house'
(2) Strictly postnominal	*l'industrie* ***chimique*** $\sim$ **la* ***chimique*** *industrie* 'chemical industry'
(3) Flexible	*une maison* ***magnifique*** $\sim$ *une* ***magnifique*** *maison* 'a beautiful house'

Table 2.1: Adjective types in French.

The Common Voice corpus of French contains 747 hours of speech, and over 130,000 noun-adjective pair tokens. Corpus results confirm that postnominal order is indeed the most common. 61% of non-unique pairs are postnominal (79613/130513; token frequency), and 69% of unique pairs are postnominal (43674/63296; type frequency). Among the flexible pairs (those that appear in both orders in the corpus), however, only 39% are postnominal-leaning (707/1813; type frequency).

The type and token frequencies of noun-adjective pairs, adjectives, and nouns found in the corpus are reported in Table 2.2. Types are unique pairs (order-sensitive), adjectives, and nouns; and tokens are all instances of pairs, adjectives, and nouns. Examining the trends at the word level, the data show that adjectives typically come after the noun they modify: 70% of unique adjectives are strictly postnominal (5360/7657; type frequency). Among the flexible adjectives, 70% are postnominal-leaning (1167/1667; type frequency). 34% of unique nouns are strictly preadjectival (3276/9635; type frequency), compared to 24% that are strictly postadjectival. Among the flexible nouns, 54% are preadjectival-leaning (2194/4063; type frequency). Additional data or different corpora may yield different frequency distributions, depending on speaker information such as dialect, or language use such as formal or informal context.

Data	Token Frequency	Type Frequency
All noun-adjective pairs	130,513	63,296
Flexible noun-adjective pairs	7,040	1,813
All adjectives	130,513	7,657
Flexible adjectives	7,040	1,667
All nouns	130,513	9,635
Flexible nouns	4,199	4,063

Table 2.2: Token and type frequencies of noun-adjective pairs, adjectives, and nouns in the Common Voice French corpus data. Flexible indicates that the pair or lexical item appeared in both positions, prenominal and postnominal.

While the flexible NP provided in Table 2.1, *magnifique* is said to have the same truth-value in both orders (Knittel, 2005), it is well known that not all adjectives or noun-adjective pairs behave in this way, as in (12) below.

(12) a. un homme grand *postnominal*
a man big
'a tall man'

b. un grand homme *prenominal*
a big man
'a great man'

The adjective *grand*, 'big' has two different senses in postnominal and prenominal position. This type of semantic effect will be handled in the statistical model using the semantic clustering method described in section 1.4.3.

2.2 Phonology

In this section, I discuss the phonological markedness constraints as they pertain to French: whether or not they can be violated at the word boundary, and if so whether they are repaired phonologically. A summary of the constraints is provided at the end of this section, in Table 2.17. The consonant and vowel inventories of the language are below, following Fougeron and Smith (1993); Tranel (1987); Dell et al. (1980). Stress in French is generally considered to be assigned only at the phrase level, making clash and lapse at the word level not possible. French syllable structure allows for complex onsets and codas.

	Bilabial	Lab. dent.	Dental	P-alveo.	Palatal	Velar
Plosive	p b		t d			k g
Nasal	m		n		ɲ	(ŋ)
Fricative		f v	s z	ʃ ʒ		ʁ
Approx				l	j	
Lab. appr.					ɥ	w

Table 2.3: Consonant inventory of French (Fougeron and Smith, 1993).

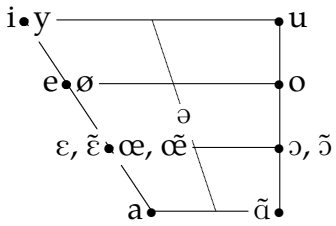


Table 2.4: Vowel inventory of French (Fougeron and Smith, 1993).

2.2.1 Vowel hiatus

Hiatus occurs when two vowels in separate syllables occur directly adjacent to one another. Hiatus can be repaired in certain environments in French through a process called *liaison*. Liaison is a phenomenon whereby final consonants that are silent in isolation, are produced between words with hiatus under certain syntactic and lexical conditions (Tranel, 1995). Syntactically, liaison can be obligatory, optional, or ungrammatical for certain structures. Between a prenominal adjective and following noun, liaison is highly frequent. In postnominal NOUN ADJECTIVE order, liaison is possible, but is considered a feature of a more "elevated" style of speaking (Morin, 2011). There are also contexts where liaison is required, such as between a definite article and noun; and where it is ungrammatical, such as between an inverted subject and verb. The conditions under which liaison surfaces are complex, and can be attributed to a number of factors including syntactic, phonological, and sociolinguistic Durand and Lyche (2008). See the examples in (13).

(13) **French liaison**

a. ADJECTIVE NOUN

grands hommes *highly frequent*
[gʁɑ̃ (**z**)ɔm]
big men
'great men'

b. NOUN ADJECTIVE

avions américains *possible, but elevated*
[avjɔ̃ (**z**)amɛʁikɛ̃]
planes american
'American planes'

c. DETERMINER NOUN

les arbres *obligatory*
[le **z**aʁbʁə]
the trees
'the trees'

d. SUBJECT PAST PARTICIPLE

Avez-vous étudié? *ungrammatical*
[avevu ∅etudje]
have-you studied
'Did you study?'

Liaison also depends on the lexical item. Some vowel-initial words prohibit liaison; orthographically they begin with an 'h,' which is "aspirated" (*h aspiré*). Examples are in (14). This set of words that tolerate hiatus is relatively small compared to the broader lexicon of French.

(14) **French *h aspiré* words**

a. grands haricots
[gʁɑ̃ ∅aʁiko]
'big beans'

b. grandes haches
[gʁɑ̃d ∅aʃ]
'big axes'

Additionally, native speakers have been observed to over-generalize the liaison process and repair hiatus at word boundaries where it does not apply in French; this phenomenon is called *pataquès*, among other names, or "false liaison" (Spitzer, 1941).

In French, instances of hiatus at the word boundary are actively avoided, as seen by the process of liaison. HIATUS is therefore predicted to have an effect on noun-adjective word ordering, given the PHONOLOGICALLY-CONDITIONED HYPOTHESIS.

Tables 2.5 and 2.6 provide descriptive statistics from the Common Voice French corpus, showing how many adjectives and nouns are vowel initial and vowel final. These data provide an idea of the phonological shape of these words, and how likely hiatus may be in noun-adjective pairs. Distributions of how often HIATUS is violated in the corpus data are in Table 2.7.

MEASURE	TOKEN		TYPE	
Vowel initial	26,997	(21%)	2,384	(31%)
Vowel final	7,386	(6%)	1,025	(13%)

Table 2.5: Proportion of adjective tokens (n=130513) and types (n=7657) that are vowel initial or vowel final in the French Common Voice corpus.

Measure	Token		Type	
Vowel initial	23,790	(18%)	2,285	(24%)
Vowel final	17,224	(13%)	1,443	(15%)

Table 2.6: Proportion of noun tokens (n=130513) and types (n=9635) that are vowel initial or vowel final in the French Common Voice corpus.

Hiatus violation	Frequency	
Prenominal order	1,113	(1%)
Postnominal order	2,742	(2%)
Both orders or neither	126,658	(97%)

Table 2.7: Violations of Hiatus among noun-adjective pairs (n=130513) in the French Common Voice corpus.

2.2.2 Voice-disagreeing clusters

Contact at the word boundary between obstruents that do not have the same specification for voicing are avoided in this constraint, Voice. In a production study, Snoeren and Segui (2003) found that French speakers produced regressive voicing assimilation between obstruents across a word boundary that did not agree in voicing. They tested stops and fricatives, in voiceless-voiced and voiced-voiceless sequences. An example is shown below, from Snoeren and Segui (2003).

(15) **Regressive voicing assimilation in French**

a. une jupe droite
[yn ʒy**b** dʁwat]
a skirt straight
'a straight skirt'

b. une robe claire
[yn ʁɔ**p** klɛʁ]
a dress light
'a light dress'

Within the word, regressive voicing assimilation has also been reported for obstruent coda clusters. In a corpus study of journalistic speech, Hallé and Adda-Decker (2007) find similar results to those reported experimentally by Snoeren and Segui (2003).

Voicing assimilation occurs at the word level and at word boundaries, including between nouns and adjectives as found in Snoeren and Segui's study. Following the PHONOLOGICALLY-CONDITIONED HYPOTHESIS, I predict that VOICE will have an effect on noun-adjective ordering.

Tables 2.8 and 2.9 provide descriptive statistics from the Common Voice French corpus, showing how many adjectives and nouns have a voiced onset, voiceless onset, voiced coda, or voiceless coda. These data provide an idea of the phonological shape of these words, and how likely a mismatch in voicing between nouns and adjectives may be. Distributions of how often VOICE is violated in the corpus data are in Table 2.10.

MEASURE	TOKEN		TYPE	
Voiceless onset	50,958	(39%)	3,354	(38%)
Voiced onset	52,558	(40%)	3,167	(36%)
Voiceless coda	24,878	(19%)	1,993	(22%)
Voiced coda	68,964	(53%)	3,508	(39%)

Table 2.8: Proportion of adjective tokens (n=130513) and types (n=7657) that have a voiced onset, voiceless onset, voiced coda, or voiceless coda in the French Common Voice corpus.

2.2.3 OCP-Place

This constraint targets consonant clusters at the word boundary that share a place of articulation. There is no strong evidence that OCP-PLACE is active in

Measure	Token	Type
Voiceless onset	58,276 (45%)	4,281 (41%)
Voiced onset	48,447 (37%)	3,817 (37%)
Voiceless coda	20,317 (16%)	1,826 (18%)
Voiced coda	50,957 (39%)	3,663 (35%)

Table 2.9: Proportion of noun tokens (n=130513) and types (n=9635) that have a voiced onset, voiceless onset, voiced coda, or voiceless coda in the French Common Voice corpus.

Voice violation	Frequency
Prenominal order	31,983 (24%)
Postnominal order	19,027 (15%)
Both orders or neither	79,503 (61%)

Table 2.10: Violations of Voice among noun-adjective pairs (n=130513) in the French Common Voice corpus.

the phonology of French, so it is expected not to be a significant predictor of noun-adjective ordering in the statistical model, following the Phonologically-conditioned Hypothesis.

In a paper describing Québec French, Côté (1997) proposes that reduction of consonant clusters in word-final position was found to be more likely if more features were shared between the consonants, place of articulation among them. The data presented in her work, however, was impressionistic and did not come from human subjects experiments or corpora, and no statistical evidence was provided for the trends or likelihoods. Additionally, cluster reduction is said to always be optional and variable across dialects; the corpus used for this dissertation is not primarily of Québec French, though it is possible that it includes some speakers of this dialect.

Tables 2.11 and 2.12 provide descriptive statistics from the Common Voice French corpus, showing how many adjectives and nouns have an onset or coda

at the attested places of articulation. These data provide an idea of the phonological shape of these words, and how likely a violation of OCP-Place between nouns and adjectives may be. Distributions of how often this constraint is violated in the corpus data are in Table 2.13.

Measure	Token		Type	
Labial onset	40,712	(31%)	2,539	(29%)
Labial coda	10,202	(8%)	561	(6%)
Coronal onset	45,633	(35%)	2,419	(27%)
Coronal coda	52,082	(40%)	3,095	(35%)
Velar onset	17,218	(13%)	1,575	(18%)
Velar coda	31,558	(24%)	1,845	(21%)

Table 2.11: Proportion of adjective tokens (n=130513) and types (n=7657) that have an onset or coda at the attested places of articulation in the French Common Voice corpus.

Measure	Token		Type	
Labial onset	41,623	(32%)	3,071	(30%)
Labial coda	8,591	(7%)	614	(6%)
Coronal onset	40,260	(31%)	2,959	(28%)
Coronal coda	37,100	(28%)	3,098	(30%)
Velar onset	24,981	(19%)	1,777	(17%)
Velar coda	25,583	(20%)	1,845	(21%)

Table 2.12: Proportion of noun tokens (n=130513) and types (n=9635) that have an onset or coda at the attested places of articulation in the French Common Voice corpus.

OCP violation	Frequency	
Prenominal order	11,939	(9%)
Postnominal order	7,718	(6%)
Both orders or neither	110,856	(85%)

Table 2.13: Violations of OCP among noun-adjective pairs (n=130513) in the French Common Voice corpus.

2.2.4 Length

This constraint is violated if the first word in a noun-adjective pair is longer than the second word, based on syllable count. Various works describe a tendency for "short before long" in noun-adjective ordering in French (Forsgren, 1978; Abeillé and Godard, 1999; Thuilier, 2012). An example is in (16), from Thuilier (2012), page 110. The disyllabic adjective *avide* 'greedy' is preferred postnominally after a monosyllabic noun, but prenominally before a polysyllabic noun.

(16) **Short before long preference**

a. un air avide
   a air greedy
   'a greedy air'

b. un avide hippopotame
   a greedy hippopotamus
   'a greedy hippopotamus'

LENGTH is predicted to have an effect on noun-adjective ordering, consistent with the previous literature.

Tables 2.14 and 2.15 provide descriptive statistics from the Common Voice French corpus, showing the mean, median, and mode lengths of adjectives and nouns by syllable count. These data provide an idea of the typical length of these words. Distributions of how often LENGTH is violated in the corpus data are in Table 2.16.

2.2.5 Stress constraints

Dell (1984) discusses the rhythm rule's application in French; however, stress

Measure	Token	Type
Mean syllable count	2.2	2.6
Median syllable count	2	3
Mode syllable count	2 (39%)	2 (36%)

Table 2.14: Mean, median, and mode syllable counts for adjectives in the French Common Voice corpus.

Measure	Token	Type
Mean syllable count	2.1	2.6
Median syllable count	2	2
Mode syllable count	2 (37%)	2 (38%)

Table 2.15: Mean, median, and mode syllable counts for nouns in the French Common Voice corpus.

Length violation	Frequency
Prenominal order	50,542 (39%)
Postnominal order	40,711 (31%)
Words are the same length	39,260 (30%)

Table 2.16: Violations of Length among noun-adjective pairs (n=130513) in the French Common Voice corpus.

assignment is generally considered to be at the phrasal level, rather than the word level. For this reason, clash and lapse are not relevant in this analysis as noun-adjective ordering is at too low of a prosodic level to be active in French.

2.3 Methods

This thesis analyzes spoken corpus data from the Common Voice corpus, provided by Mozilla[2]. The French data analyzed in this work come from version fr_834h_2021-07-21, consisting of 747 hours of validated speech from 15,391

[2]Accessed Fall 2021 at the following address: voice.mozilla.org

Constraint	Active status
Clash	Not possible
Lapse	Not possible
Hiatus	Active across word boundaries (*liaison*; (Tranel, 1995))
Voice	Active across word boundaries (*regressive assimilation*;, (Snoeren and Segui, 2003))
OCP-Place	Not active
Length	Active for noun-adjective pairs (Forsgren, 1978; Thuilier, 2012)

Table 2.17: Summary table of which phonological constraints are active in French.

speakers. Dialect information of the speakers was limited: 63% of data comes from France, 2% from Canada, 2% from Belgium, and 1% from Switzerland, the remaining 32% is not reported. Noun-adjective pairs were extracted after the sentences were tagged using `spaCy`, whose models have a 96% accuracy on part-of-speech tagging for French[3].

The majority of the analysis was carried out on the phonemic representations of the audio in Common Voice. Common Voice is transcribed orthographically, and was converted to the phonemic level using lexical databases. The lexical database for French comes from Lexique 3 (New et al., 2004), and consists of 140,000 word forms. 28% of the data were excluded due to missing pronunciations of one or both members of the noun-adjective pair (14513/51830)[4]. The phonological information in this database includes phonemes, syllable boundaries, and liaison context pronunciation, all of which were used to code the con-

[3]Models were trained and tested on data from the Universal Dependencies French Sequoia corpus (Candito and Seddah, 2012), the named-entity recognition Wikipedia corpus (Nothman et al., 2017), and `spaCy` lookups data, located at: github.com/explosion/spacy-lookups-data. In general, `spaCy` model accuracy was likely evaluated on held out data from Wikipedia, text, and formal speech.

[4]The exclusion of so much data is not ideal, and while a more complete dataset may affect results, the overall analysis is expected to remain the same as there were no known patterns characterizing datapoints that were excluded.

straints analyzed here. Specific constraint definitions are described in Table 2.18.

Constraint	Definition
Clash	Not possible for French.
Lapse	Not possible for French.
Hiatus	Two members of the vowel set adjacent at the word boundary: {i y e ø ɛ œ a u o ɔ ɛ̃ œ̃ ɔ̃ ɑ̃}. Words with a liaison consonant do not violate hiatus, except before words with an *h aspiré*[5].
Voice	One voiceless consonant and one voiced consonant adjacent at the word boundary: {p t k f s ʃ}, {b d g v z ʒ m n ɲ ŋ l ʁ j ɥ w}
OCP-Place	Two consonants at the same place of articulation adjacent at the word boundary: {p b f v w ɥ}, {t d s z l}, {ʃ ʒ j ɥ}, {k g ʁ w}

Table 2.18: Definition of phonological constraints for French.

The corpus data were split into two groups based on semantic similarity between the embeddings of the adjective in its prenominal and postnominal positions. The distribution of cosine similarities, shown in Figure 2.1, was fit for two distributions using a Gaussian mixture model. The boundary between these two distributions, marked with a red vertical line in the figure, was used to bin the data.

Pair tokens containing an adjective with a cosine similarity below 0.47 were categorized in the *dissimilar* dataset, and those with a similarity above 0.47 were categorized in the *similar* dataset. This threshold was verified with manual inspection of 22 adjectives listed in *Le Petit Robert* with a specific meaning for at least one of the positions relative to the noun (usually prenominal); only two of them have cosine similarity values that put them above the threshold, incorrectly in the *similar* dataset. This is shown in Table 2.19. Some noise was expected as

[5]This is a conservative estimate, as no distinction is made for register (see 2.2.1 more details on French liaison). All potential liaison instances are treated as surfacing, meaning vowel hiatus occurs potentially more rarely in this dataset than in actual speech.

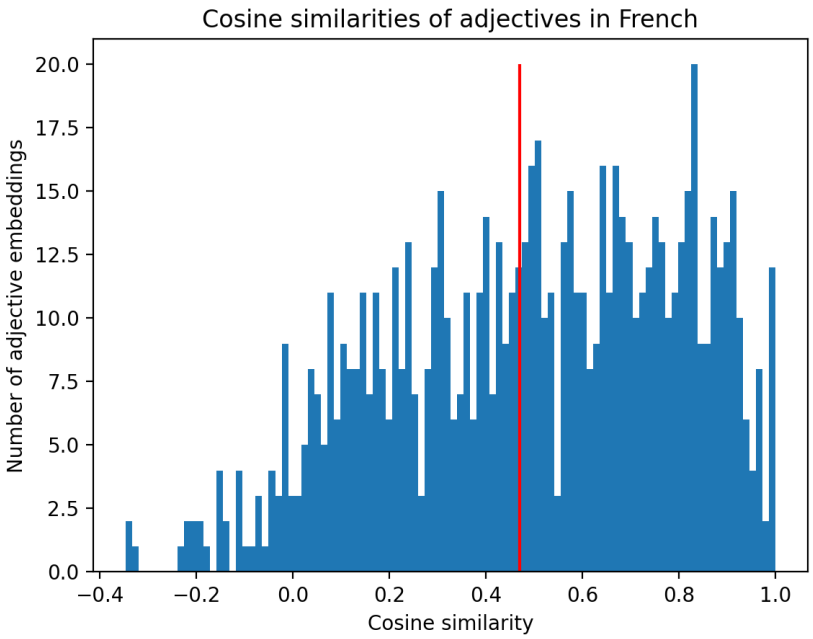


Figure 2.1: Distribution of cosine similarities measured between prenominal and postnominal embeddings of adjectives in French. Cut-off of 0.47 is marked with a red vertical line.

this is an automated measure.

A sample of 50 tokens in which voice-disagreeing clusters are tolerated at the phonemic level between a noun and a flexible adjective were examined in Praat. These tokens all had a relative frequency between 0.5 and 0.7, meaning they ranged from appearing equally in both prenominal and postnominal order, to appearing in one order 70% of the time. Pairs were eliminated based on: clarity of the recording, perceived nativeness of the speaker, and bias for a diversity of pair types. I, a near-native speaker, analyzed the 50 recordings in Praat (Boersma and Weenik (2022)), to judge if any phonological repairs for voice-disagreeing clusters in French were present in the tokens, namely regressive voicing assimilation (see section 2.2.2). Voicing was determined by the presence or absence of a voicing bar in the spectrogram.

Dissimilar adjectives		Cosine similarity
ancien	'former/old'	0.07
certain	'certain'	0.34
cher	**'dear/expensive'**	**0.80**
curieux	'curious'	0.32
dernier	'last'	0.02
prochain	'next'	0.25
différent	'different'	0.23
divers	**'multiple different/assorted'**	**0.75**
fameux	'first rate/famous'	0.03
franc	'hearty/clean'	-0.03
grand	'great/tall'	0.22
même	'same'	0.17
nouveau	'new'	0.08
pauvre	'pitiful/poor'	0.20
premier	'primary'	0.17
propre	'own/clean'	0.16
pur	'simple/fresh'	0.22
rare	'extraordinary/uncommon'	0.21
seul	'only/alone'	0.06
simple	'simple/modest'	0.19
unique	'single/only'	0.32
vrai	'real/true'	0.26

Table 2.19: Cosine similarity values for adjectives with position-specific definitions in *Le Petit Robert*. Those above the threshold of 0.47 are in bold, showing that they are incorrectly categorized.

2.4 Results

2.4.1 Regression models

A mixed effects logistic regression was fit to the French corpus data using `glmer` in R (R Core Team, 2016). The model predicted the order of pair tokens (prenominal ADJECTIVE NOUN or postnominal NOUN ADJECTIVE), using the phonological constraints possible at the word boundary in French: HIATUS, VOICE, OCP, and LENGTH, in addition to RELATIVE FREQUENCY. For more information about

how phonological constraints were coded, see 1.4.2. The models also included random intercepts for Adjective and Noun lemmas[6]. See Figure 2.2 at the end of this section for a visualization of the results from both models.

$$\texttt{glmer}(\text{outcome} \sim \text{Hiatus} + \text{Voice} + \text{OCP} + \text{Length} + \text{Relative frequency} + (1|\,\text{ADJ}) + (1|\,\text{NOUN}),\ family = \text{“binomial”}\ )$$

Regression models were fit separately for the two datasets separated by the semantic difference threshold, with the prediction that the phonological effects would be relatively stronger, or present only in the *more similar* dataset which contains adjectives with a higher cosine similarity between their prenominal and postnominal embeddings. Results from the *dissimilar* model are presented in Table 2.20, and those from the *similar* model in Table 2.21.

	Estimate	Std. Error	z value	p value
Intercept	**-5.23076**	**0.23834**	**-21.947**	**< 2e-16 *****
Constraint: Hiatus (-1)	**-1.75777**	**0.30737**	**-5.719**	**1.07e-08 *****
Constraint: Hiatus (1)	**0.63725**	**0.13955**	**4.566**	**4.96e-06 *****
Constraint: Voice (-1)	**-0.21086**	**0.05230**	**-4.032**	**5.53e-05 *****
Constraint: Voice (1)	-0.01367	0.05997	-0.228	0.8197
Constraint: OCP (-1)	-0.13982	0.07640	-1.830	0.0672
Constraint: OCP (1)	0.02813	0.08214	0.343	0.7320
Constraint: Length (-1)	0.10420	0.05848	1.782	0.0748
Constraint: Length (1)	0.06237	0.05620	1.110	0.2670
Relative frequency	**3.34314**	**0.18410**	**18.159**	**< 2e-16 *****

Table 2.20: Model fit for French data containing *less similar* adjectives. Number of observations is 61,060 noun-adjective pairs. R^2 = 0.41.

In the *dissimilar* dataset, Hiatus, Voice, and Relative frequency are all significant predictors of order. Hiatus (-1) has a negative coefficient, indicating that

[6]Random slopes were not included because the increase in the complexity caused the model to take too long to fit.

a noun-adjective pair with a violation of the constraint in prenominal order but not postnominal order, is more likely to be postnominal compared to when the constraint is inactive (i.e., is violated in both orders or in neither order, which are both coded as zero). Showing a complementary effect, HIATUS (1) has a positive coefficient, indicating that a pair with a violation of the constraint in postnominal order but not prenominal order is likelier to be prenominal compared to when the constraint is inactive. VOICE (-1) has a negative coefficient, indicating that a violation of the constraint in prenominal order, but not in postnominal order, correlates with postnominal order. RELATIVE FREQUENCY has a positive coefficient indicating that pairs with a greater degree of flexibility are prenominal-leaning, meaning the likelihood of a pair surfacing as prenominal increases as its flexibility increases.

	ESTIMATE	STD. ERROR	Z VALUE	P VALUE
Intercept	0.59863	0.37345	1.603	0.10894
Constraint: HIATUS (-1)	**-1.24683**	**0.29925**	**-4.167**	**3.09e-05 *** **
Constraint: HIATUS (1)	**0.62272**	**0.20029**	**3.109**	**0.00188 ** **
Constraint: VOICE (-1)	-0.09894	0.06824	-1.450	0.14713
Constraint: VOICE (1)	-0.15396	0.08690	-1.772	0.07644
Constraint: OCP (-1)	-0.07879	0.09922	-0.794	0.42712
Constraint: OCP (1)	0.22890	0.11723	1.953	0.05087
Constraint: LENGTH (-1)	-0.12001	0.08055	-1.490	0.13625
Constraint: LENGTH (1)	-0.07603	0.07399	-1.028	0.30413
RELATIVE FREQUENCY	**-1.73511**	**0.35287**	**-4.917**	**8.78e-07 *** **

Table 2.21: Model fit for French data containing *more similar* adjectives. Number of observations is 34,259 noun-adjective pair tokens. $R^2 = 0.55$.

In the *similar* dataset, HIATUS and RELATIVE FREQUENCY are significant predictors of order. HIATUS (-1) has a negative coefficient and HIATUS (1) has a positive coefficient, indicating that a noun-adjective pair with a violation of the constraint in one order but not the other, is likelier to be in the order that avoids that vi-

olation compared to when the constraint is inactive. RELATIVE FREQUENCY has a negative coefficient, indicating that pairs with a greater degree of flexibility are postnominal-leaning, meaning the likelihood of a pair surfacing as postnominal increases as its flexibility increases.

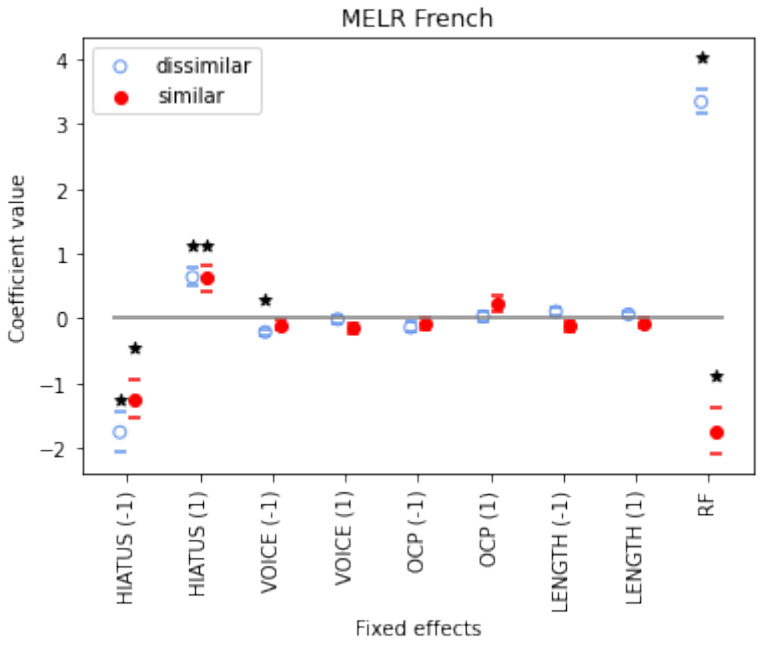


Figure 2.2: Visualization of the mixed-effects logistic regression results in French of the *dissimilar* (blue, triangle) and *similar* (red, circle) models. Coefficient values with standard error are shown for each fixed effect; significant effects are indicated by a star.

One-tailed Z-tests comparing HIATUS (-1) across models and HIATUS (1) across models were not run because the absolute values of the coefficients were not greater in the semantically similar model than in the dissimilar model, as hypothesized.

2.4.2 Acoustic Sample

50 tokens (41 types) of flexible noun-adjective pairs that have underlying voice-disagreeing clusters at the word boundary were examined at the phonetic level. 35/50 of the tokens had some type of repair, while the remaining 15 tokens did not have any repair – the disagreement in voicing in cluster was produced in the token. Most of the repairs were cases of regressive voicing assimilation (25/35), whereby the coda consonant of the first word was produced with the same voicing as the onset of the second word. There were 19 cases of full assimilation and six cases of partial assimilation. See an example in Figure 2.3 of full assimilation in the noun-adjective pair *nombreuses fleurs*, 'many flowers.' In this example, the final consonant in *nombreuses* /nɔ̃.bʁøz/ is expected to be underlyingly voiced, but is produced as voiceless before *fleurs* /flœʁ/, which has an initial voiceless consonant. Partial assimilation can be seen in Figure 2.4, of the final voiced consonant in noun *rédacteur* /ʁe.dak.tœʁ/, coming before the initial voiceless consonant in *principal* /pʁɛ̃.si.pal/, in the pair *rédacteur principal* 'main editor.'

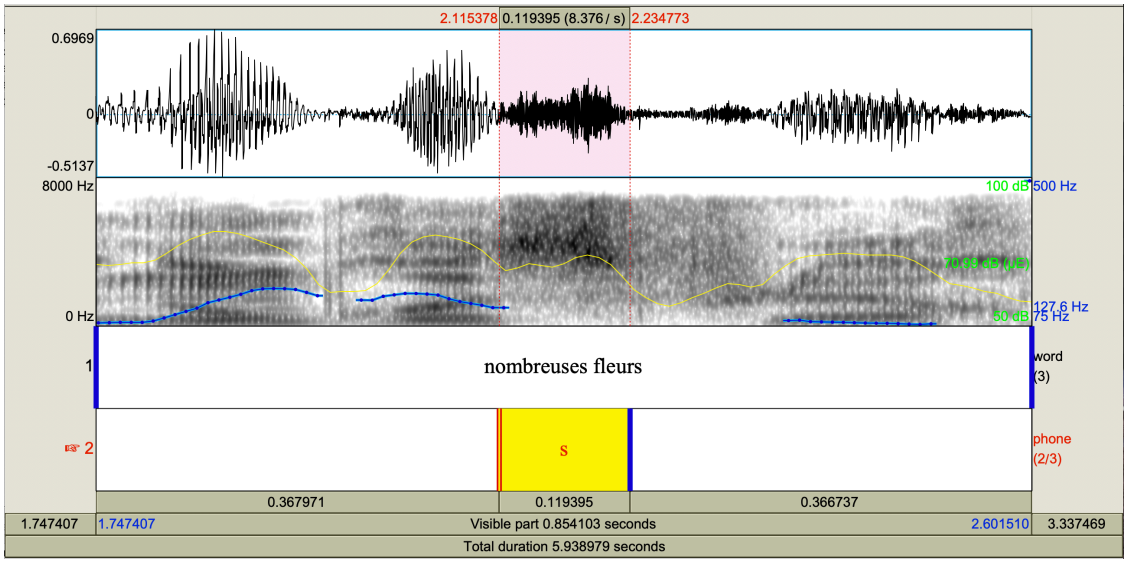


Figure 2.3: Full assimilation of /z/ to [s] in the phrase *nombreuses fleurs* 'many flowers', spoken by a male French speaker in the Common Voice corpus (common_voice_fr_20269979.mp3). Pitch is tracked in blue and intensity in yellow in the spectrogram.

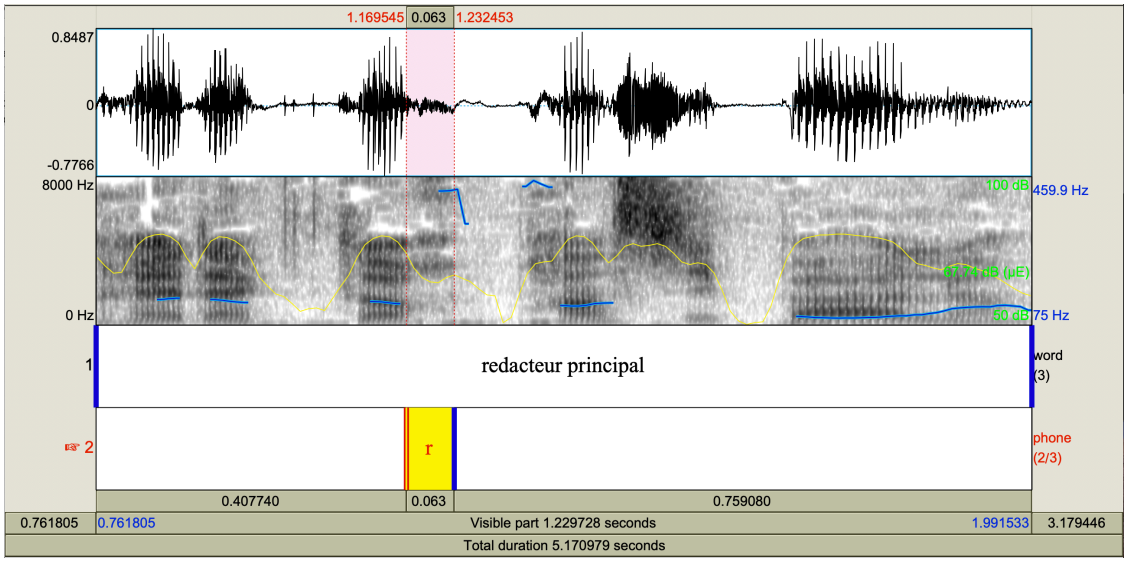


Figure 2.4: Partial assimilation of /ʁ/ to [χ] in the phrase *rédacteur principal* 'main editor', spoken by a male French speaker in the Common Voice corpus (common_voice_fr_19812631.mp3). Pitch is tracked in blue and intensity in yellow in the spectrogram.

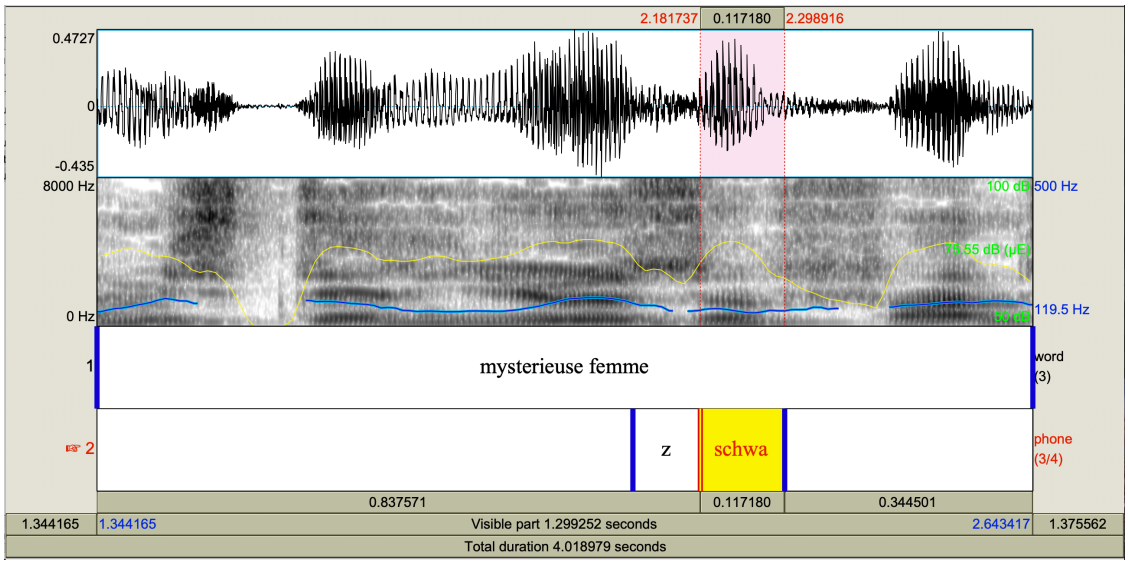


Figure 2.5: Pronunciation of an *e muet* in *mystérieuse femme* 'mysterious woman', spoken by a female French speaker in the Common Voice corpus (common_voice_fr_23892094.mp3). Pitch is tracked in blue and intensity in yellow in the spectrogram.

There were ten cases where an optional final schwa was produced, breaking up the cluster at the word boundary that disagreed in voicing. Such a schwa is referred to as an *e muet*, and the conditions surrounding the probability of this vowel surfacing are complex (e.g., Lucci, 1976; Berri, 2006; Griffiths et al., 2020). The interaction between the voicing of a consonant cluster across a word boundary and the likelihood of a word final *e muet* is an avenue for future research. An example of this phenomenon is shown in Figure 2.5. A final schwa is produced between *mystérieuse* /mi.ste.ʁjøz/ and *femme* /fam/, effectively eliminating the surfacing of a cluster that disagrees in voicing in the pair *mystérieuse femme* 'mysterious woman.'

Finally, the remaining 15 tokens had no repair: the cluster was produced with a disagreement in voicing. This is shown in Figure 2.6, where the final consonant in *proche* /pʁɔʃ/ surfaces as voiceless before the initial voiced consonant in *vallée*

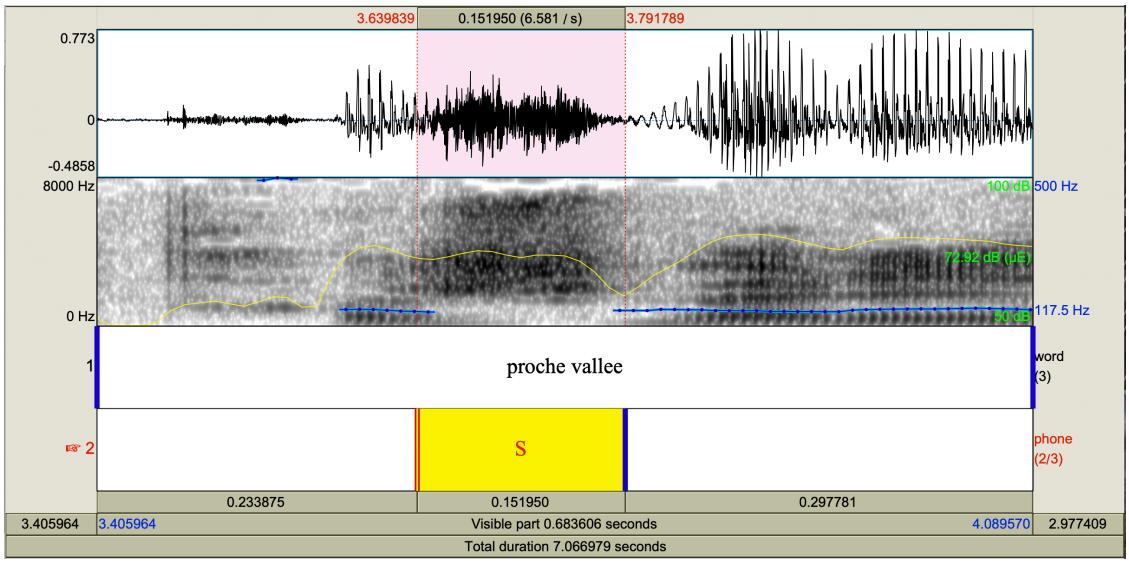


Figure 2.6: No assimilation of /ʃ/ in the phrase *proche vallée* 'nearby valley', spoken by a male French speaker in the Common Voice corpus (common_voice_fr_19781758.mp3). Pitch is tracked in blue and intensity in yellow in the spectrogram.

Surface phenomenon	Frequency
Full assimilation	19 (38%)
Partial assimilation	6 (12%)
e muet	10 (20%)
None	15 (30%)
Total	50

Table 2.22: Results summary for the acoustic analysis of the sample of tolerated voice-disagreeing clusters in French.

/va.le/ in the pair *proche vallée* 'nearby valley.'

Table 2.22 provides a summary of the results found in the acoustic sample of tolerated Voice in French. Among the cases where no assimilation was found, two of them where instances where *e muet* was also not possible (i.e., the lexical item that was word 1 could not be produced with a final schwa).

2.5 Discussion

2.5.1 Regression models

Following the PHONOLOGICALLY-CONDITIONED HYPOTHESIS, the constraints HIATUS, VOICE, and LENGTH were predicted to have significantly positive effects on noun-adjective word ordering, where the adjective is flexible. The evidence presented in section 2.2 supports the claim that these three constraints are active in French, at least at the word level, if not also between words; therefore, they were predicted to have an effect on noun-adjective ordering. These three effects were predicted to be significantly positive in at least the *more similar* regression model, or to have a larger coefficient in the *more similar* regression model compared to the *less similar* model, if significant in both, following the SEMANTICALLY-CONDITIONED HYPOTHESIS.

These phonological predictions were true for HIATUS and VOICE. Both HIATUS coefficients are significant in the two models, and VOICE (-1) is a significant predictor in the dissimilar model. These results indicate that sequences of two vowels or consonants that disagree in voicing are likely to be avoided where possible. Given the productivity of *liaison*, I expected a relatively large and consistent effect of avoiding hiatus via word ordering in the language.

LENGTH was expected to be a significant predictor, but the results from both models show that it does not have an effect on ordering. This lack of effect could be due to its previous descriptions in the literature as an observed "tendency" that is frequently violated (Forsgren, 1978; Abeillé and Godard, 1999; Thuilier, 2012).

Unexpectedly, VOICE is a relatively small but significant predictor in the *dissimilar* model, but not in the *similar* model. While there are more datapoints in the *dissimilar* model, it is also possible that avoidance of voiced-voiceless or voiceless-voiced consonant sequences at the word boundary can be attributed to a different strategy than word ordering. At the lexical level, speakers may be selecting a semantically-similar word that does not violate VOICE, rather than using word-order manipulation. Synonym selection was found to have a significant effect on phrase formation in English (Breiss and Hayes, 2020; Schlüter and Knappe, 2018), and may also play a role in French noun-adjective ordering. The implications of such effects extend the scope of phonological conditioning beyond word ordering to lexical selection (Schlüter and Knappe, 2018). The significance of synonym selection on the avoidance of VOICE violations is left to future work[7]. Another possibility is that VOICE can be repaired phonologically at all word boundaries (i.e., there are no restrictions on regressive voicing assimilation, Snoeren and Segui (2003)); whereas, liaison is only possible at some boundaries, and only with a subset of the lexicon, to repair HIATUS (Tranel, 1995). That is to say, the difference in effects between VOICE and HIATUS may be attributable to the fact that VOICE can be repaired without word-order manipulation, whereas in many cases HIATUS cannot.

As predicted, OCP is not a significant predictor of order, given the very little evidence for the avoidance of consonant sequences at the same place of articulation in French.

The semantic prediction that constraints be significantly positive or have a

[7]The strategies of word ordering versus lexical selection may be better examined in a controlled experiment, where stimuli can be designed specifically for their syntactic, semantic, and phonological attributes, as has been done in previous work (Schlüter, 2005; Schlüter and Knappe, 2018).

larger coefficient in the similar regression model compared to the dissimilar model, was not found in the data. VOICE (-1) was significant only in the dissimilar model; and, the HIATUS effects were not greater in the similar model, so no Z-test was run.

RELATIVE FREQUENCY is a significant predictor in both models; however, this coefficient is positive in the *less similar* model and negative in the *more similar* model. This indicates that as the semantic difference between an adjective in prenominal position versus postnominal position is minimized, flexible noun-adjective pairs are likelier to occur in postnominal order, which is the default in French. Inversely, if there is a greater semantic difference between positions of an adjective, flexible pairs are likelier to occur in prenominal order, which is the position to which specific or special meanings of an adjective are usually attributed (Nølke, 1996; Laenzlinger, 2005; Cinque, 2010).

2.5.2 Acoustic sample

As discussed in the previous section, one hypothesis for the non-avoidance effect of VOICE in French is that phonological repairs are preferred to different linear orderings, where possible. Results from the acoustic sample are encouraging: 70% of pairs in the sample had a phonological repair that prevented the voice-disagreeing cluster from surfacing. Most of the time, speakers did not produce a violation of VOICE at the phonetic level. This may be a case where the violation of a phonological constraint is not substantial enough to constitute the surfacing of an alternative word ordering. Such an interaction is discussed more in section 5.2.4.

CHAPTER 3

ITALIAN

Italian is a Romance language (Indo-European), spoken primarily in Italy and the European Union, with an estimated 85 million speakers worldwide. This chapter mainly discusses aspects of Standard Italian, a general term for the variety of Italian widely spoken in Italy by the educated population (Berruto, 1987). The language situation in Italy is such that Standard Italian is spoken alongside minority Romance languages (often referred to as *dialetti* 'dialects'), causing there to often be large variation between geographic regions in terms of local language and dialect or accent of the standard. The data from the Common Voice Italian corpus has no dialect information, and contains speech from over 6,000 speakers. Given that there is such a strong presence of dialects in Italy, a corpus that has enough dialect information to run dialect-specific analysis may yield clearer results.

The canonical constituent order of Italian is Subject-Verb-Object (SVO), but other orders are acceptable (Maiden and Robustelli, 2014). Especially relevant to this work is the ordering of adjective and noun, which is canonically NOUN ADJECTIVE (postnominal), but prenominal order is allowed for some adjectives and required for others. This is further elaborated on in the next section. Section 3.2 discusses how each phonological constraint is treated in Italian phonology, and includes descriptive statistics and predictions about the corpus data. Section 3.3 provides details on methodology specific to the Italian analysis. Results are presented in section 3.4, and discussed in section 3.5.

3.1 Noun-adjective flexibility

The canonical ordering of adjectives with respect to the noun that they modify is postnominal: NOUN ADJECTIVE. There are some adjectives that can occur only in prenominal position, and some than can occur both before and after the noun. Prenominal order has been described as expressing emphasis, and qualities that are generic, habitual, or essential (Hall, 1948). Adjectives that can appear in both positions may have a difference in meaning, as described by Cinque (2010) (see Table 1.1). Thus, adjectives in Italian can be thought of as belonging to one of three groups: strictly postnominal, strictly prenominal, or able to appear in both positions. Examples of all three appear in Table 3.1 (data from Cinque (2010)).

TYPE	ITALIAN
(1) Strictly prenominal	**uno ritardo* ***mero*** *un* ***mero*** *ritardo* 'a mere delay'
(2) Strictly postnominal	*un ingegnere* ***elettronico*** **un* ***elettronico*** *ingegnere* 'an electrical engineer'
(3) Flexible	*il contributo* ***prezioso*** *il* ***prezioso*** *contributo* 'precious contribution'

Table 3.1: Adjective types in Italian.

The Common Voice corpus of Italian contains 288 hours of speech, and about 73,000 noun-adjective pair tokens. It contains speech from speakers with a variety of accents. Corpus results confirm that postnominal order is indeed the most common. 60% of pairs are postnominal (43705/72841; token frequency), and 64% of unique pairs are postnominal (24531/38329; type frequency). Among the flexible pairs (those that appear in both orders in the corpus), only 48% are postnominal-leaning (673/1402; type frequency).

The type and token frequencies of noun-adjective pairs, adjectives, and nouns found in the corpus are reported in Table 3.2. Examining the trends at the word level, the data show that adjectives typically come after the noun they modify: 67% of unique adjectives are strictly postnominal (3756/5606; type frequency). Among the flexible adjectives, 59% are postnominal-leaning (727/1233; type frequency). Nouns typically precede their modifiers: 66% of unique nouns are strictly preadjectival (4319/6544; type frequency). Among the flexible nouns, 54% are preadjectival-leaning (1554/2878; type frequency).

Data	Token Frequency	Type Frequency
All noun-adjective pairs	72,841	38,329
Flexible noun-adjective pairs	4,199	1,402
All adjectives	72,841	8,705
Flexible adjectives	4,199	1,233
All nouns	72,841	8,715
Flexible nouns	4,199	2,878

Table 3.2: Token and type frequencies of noun-adjective pairs, adjectives, and nouns in the Common Voice Italian corpus data. Flexible indicates that the pair or lexical item appeared in both orders, PRENOMINAL and POSTNOMINAL.

While the flexible NP provided in Table 3.1 is said to have the same truth-value in both orders (Cinque, 2010), it is well known that not all adjectives or noun-adjective pairs behave in this way. See example (17) below.

(17) a. un uomo povero *postnominal*
a man poor
'a poor man' (not rich)

b. un povero uomo *prenominal*
a poor man
'a pitiful man'

The adjective *povero*, 'poor' has two different senses in postnominal and

prenominal position. This type of semantic effect will be handled in the statistical model using the semantic clustering method described in section 1.4.3.

3.2 Phonology

In this section, I discuss the phonological markedness constraints as they pertain to Italian: whether or not they can be violated at the word boundary, and if so whether they are repaired phonologically. A summary of the constraints is provided at the end of this section, in Table 3.17. The consonant and vowel inventories of the language are below, following Hall (1948); Kramer (2009). Stress in Italian is most frequently on the penultimate syllable, but other positions are possible, including final and initial which are relevant for stress clash. While coda consonants are possible word-internally, word-finally, codas are extremely rare and occur mostly in loan words.

	Bilabial	Lab. dent.	Dental	Alveolar	P-alveo.	Palatal	Velar
Plosive	p b		t d				k g
Affricate			ts dz		ʧ ʤ		
Nasal	m			n		ɲ	
Trill				r			
Fricative		f v	s z		ʃ		
Approx						j	w
Lat. appr.				l		ʎ	

Table 3.3: Consonant inventory of Italian (Kramer, 2009).

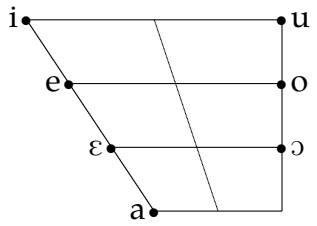


Table 3.4: Vowel inventory of Italian (Kramer, 2009).

3.2.1 Stress clash

Stress clash occurs when two stressed syllables are adjacent to each other. Clash is possible in Italian, occurring between words that have final and initial stress. This is shown in (18), where stressed syllables are underlined. In canonical post-nominal order, stress clash occurs between the noun and adjective, but is avoided in the use of prenominal order.

(18) **Italian clash**

a. città vecchia — *postnominal clash*
city old

b. vecchia città — *prenominal avoided clash*
old city
'old city'

Clash has been said to be repaired phonologically in Italian with stress retraction or a process called *Raddoppiamento Sintattico* (RS). Nespor and Vogel (1979) find that speakers retract stress from word 1 back a syllable when word 2 had stress on the first syllable (i.e., there was a stress clash at the boundary between word 1 and word 2, as in 18a). Results showed this was more likely to occur when word 1 and word 2 belonged to the same syntactic phrase, such as in the present work which investigates noun-adjective pairs.

In the same study, Nespor and Vogel (1979) found that some speakers showed initial lengthening of the onset consonant in word 2 in a stress clash environment, referred to as RS, rather than stress retraction. An example is shown in (19). The geminate onset, as opposed to the underlying singleton onset, creates distance between the two adjacent stressed syllables. While this has traditionally been described as a repair for clash, this gemination phenomenon may in fact be the result of the weight to stress principle (Prince, 1990), in which case it would be a repair of syllable organization rather than prominence (Burroni, 2022). The exact nature of RS is left to future work.

(19) **Italian *raddoppiamento sintattico***

a. città **vv**ecchia — *application of RS*
city old

b. vecchia città — *RS not needed*
old city
'old city'

Nespor and Vogel (1979) found that speakers employ either stress retraction or RS, but never both. In an attempt to increase distance between the two prominent syllables if stress clash must occur, stress retraction varieties do also exhibit lengthening, but of the final vowel in word 1. This is often referred to as the rhythm rule.

Italian actively avoids instances of clash at the word boundary, as seen by the processes of retraction or RS[1]. Clash is therefore predicted to have an effect on noun-adjective word ordering, following the Phonologically-conditioned Hypothesis.

[1]This conclusion builds on previous literature, as described in this section. The re-analysis of clash avoidance strategies as consequences of the weight-to-stress principle (Prince, 1983; Prince and Smolensky, 2004) is an issue left to future work.

Tables 3.5 and 3.6 provide descriptive statistics from the Common Voice Italian corpus, showing how many adjectives and nouns have initial or final stress, as well as the number of monosyllabic words. These data provide an idea of the phonological shape of these words, and how likely clash may be in noun-adjective pairs. Distributions of how often CLASH is violated in the corpus data are in Table 3.7.

MEASURE	TOKEN		TYPE	
Initial stress	25,270	(35%)	1,563	(18%)
Final stress	953	(1%)	83	(1%)
Monosyllables	1,001	(1%)	45	(1%)

Table 3.5: Proportion of adjective tokens (n= 72841) and types (n=8705) that have initial, final, or penultimate stress in the Italian Common Voice corpus.

MEASURE	TOKEN		TYPE	
Initial stress	28,852	(40%)	2,129	(24%)
Final stress	2,810	(4%)	260	(3%)
Monosyllables	564	(1%)	70	(1%)

Table 3.6: Proportion of noun tokens (n=72841) and types (n=8715) that have initial, final, or penultimate stress in the Italian Common Voice corpus.

CLASH VIOLATION	FREQUENCY	
Prenominal order	1079	(1.5%)
Postnominal order	1068	(1.5%)
Both orders or neither	70694	(97%)

Table 3.7: Violations of CLASH among noun-adjective pairs (n=72841) in the Italian Common Voice corpus.

3.2.2 Stress lapse

Stress lapse occurs when three or more unstressed syllables are adjacent. Lapse is also said to be dispreferred in Italian, though not to the same extent as has been reported for clash (Nespor and Vogel, 1989). Phonologically, it is repaired with the addition of prominence to a weak syllable in a string of at least three unstressed syllables. Like clash, lapse also operates at the level of the phonological word, and is therefore expected to have an effect on noun-adjective ordering as nouns and adjectives constitute separate phonological words (Nespor, 2019). An example is shown below, where lapse occurs in (20a) in prenominal order, but is avoided in postnominal order in (20b).

(20) **Italian lapse**

a. vecchia alleanza *prenominal lapse*
old alliance

b. alleanza vecchia *postnominal avoided lapse*
alliance old
'old alliance'

Italian actively avoids instances of lapse across words, as seen by the process of beat addition, shown in (21). Lapse is therefore predicted to have an effect on noun-adjective word ordering, following the Phonologically-conditioned Hypothesis.

(21) **Beat addition**

a. Stefano se ne va *instance of lapse*
Stefano se ne va *beat addition on* se
'Stefano goes away.' Nespor and Vogel (1989)

Tables 3.8 and 3.9 provide descriptive statistics from the Common Voice Italian corpus, showing how many adjectives and nouns have antepenultimate or peninitial stress (the combination of which would lead to a lapse); and penultimate stress and stress on the third syllable (which would also lead to a lapse). These data provide an idea of the phonological shape of these words, and how likely lapse may be in noun-adjective pairs. Distributions of how often Lapse is violated in the corpus data are in Table 3.10.

Measure	Token		Type	
Antepenultimate stress	14,050	(19%)	2,133	(25%)
Peninitial stress	21,546	(30%)	2,978	(34%)
Penultimate stress	38,068	(52%)	5,467	(63%)
Third syllable stress	18,048	(25%)	2,764	(32%)

Table 3.8: Proportion of adjective tokens (n=72841) and types (n=8705) that have initial, final, or penultimate stress in the Italian Common Voice corpus.

Measure	Token		Type	
Antepenultimate stress	7,008	(10%)	989	(11%)
Peninitial stress	23,698	(33%)	2,921	(34%)
Penultimate stress	38,039	(52%)	5,766	(66%)
Third syllable stress	12,749	(18%)	2,140	(25%)

Table 3.9: Proportion of noun tokens (n=72841) and types (n=8715) that have initial, final, or penultimate stress in the Italian Common Voice corpus.

Lapse violation	Frequency	
Prenominal order	16,158	(22%)
Postnominal order	18,069	(25%)
Both orders or neither	38,614	(53%)

Table 3.10: Violations of Lapse among noun-adjective pairs (n=72841) in the Italian Common Voice corpus.

3.2.3 Vowel hiatus

Hiatus occurs when two vowels are adjacent to each other, and they belong to separate syllables. Word-internal examples from Italian are shown in (22) below, where in (22a) the high vowel becomes a glide, but nothing is done to repair hiatus in (22b) (data from Kramer, 2009, p.52). An epenthetic consonant appears in the fixed expression 'and here,' shown in (22c); however, epenthesis is not productive in the language. Vowel elision occurs between vowel-final articles and vowel-initial nouns, shown in (22d).

(22) a. buono
[ˈbwɔ.no]
'good'

b. paura
[pa.ˈuː.ɾa]
'fear'

c. ed ecco
[ed ek.ko]
'and here'

d. la università → l'università
[la u.ni.ver.si.ta]
'the university'

This relatively mixed tolerance-level of hiatus is in stark contrast to French, which has quite a complex and pervasive vowel hiatus repair phenomenon, liaison (see section 2.2.1). In Italian, epenthetic consonants appear in some fixed expressions to avoid it, and it is avoided via vowel deletion between clitics and nouns. The clitic-noun relationship could be a case of exceptionality due to the particular dependence of clitics on nouns, or weak faithfulness; but, vowel hiatus is often permitted as exemplified by (22b) above, given the right circumstances[2].

[2]Such as lexical stress assignment to the second vowel in a V.V sequence, and the particular

Because of this, results for syntactic avoidance of hiatus in this study are expected to be mixed as well. This is in contrast to the predictions for stress clash. Since clash has been shown to be active in Italian phonology, it is predicted to be actively avoided by means of word-order manipulation.

Given that there is no strong evidence that HIATUS is active in the phonology of Italian, it is expected not to be a significant predictor of noun-adjective ordering in the statistical model, following the predictions of the PHONOLOGICALLY-CONDITIONED HYPOTHESIS.

Tables 3.11 and 3.12 provide descriptive statistics from the Common Voice Italian corpus, showing how many adjectives and nouns are vowel initial and vowel final. These data provide an idea of the phonological shape of these words, and how likely hiatus may be in noun-adjective pairs. Distributions of how often HIATUS is violated in the corpus data are in Table 3.13.

MEASURE	TOKEN		TYPE	
Vowel initial	12,937	(18%)	2,170	(25%)
Vowel final	71,235	(98%)	8,644	(99%)

Table 3.11: Proportion of adjective tokens (n=72841) and types (n=8705) that are vowel initial or vowel final in the Italian Common Voice corpus.

MEASURE	TOKEN		TYPE	
Vowel initial	8,886	(12%)	1,570	(18%)
Vowel final	71,916	(99%)	8,585	(99%)

Table 3.12: Proportion of noun tokens (n=72841) and types (n=8715) that are vowel initial or vowel final in the Italian Common Voice corpus.

quality of the two vowels. See Kramer (2009) for more details.

HIATUS VIOLATION	FREQUENCY
Prenominal order	7,408 (10%)
Postnominal order	11,426 (16%)
Both orders or neither	54,007 (74%)

Table 3.13: Violations of HIATUS among noun-adjective pairs (n=72841) in the Italian Common Voice corpus.

3.2.4 Length

Italian is typically described as an SVO language, but other sentence structures are possible and exploited for emphasis or artistic motivations. For example, the sentence in (23a) is in the unmarked order, SVO; however, the OVS order shown in (23b) is also grammatical. This OVS structure is ungrammatical, however, when the subject NP is light (only one phonological word). This is shown in (23c). Data come from Cardinaletti (2010).

(23) a. Il partito di maggioranza fece poi la stessa proposta.
the party of majority made then the same proposal

b. **La stessa proposta** fece poi il partito di maggioranza.
the same proposal made then the party of majority
'The majority party then made the same proposal (not a similar one)'

c. ***La stessa proposta** fece poi Gianni/lui.

Given this process of Heavy NP shift in Italian, here involving heavy subjects, it is predicted more generally that the structure within a constituent such as the noun-adjective pair, is sensitive to the weight of a NP, just as it is in the case of SVO → OVS word order. LENGTH is predicted to have an effect on noun-adjective ordering, given the sensitivity to weight of other structures in the language.

Tables 3.14 and 3.15 provide descriptive statistics from the Common Voice

Italian corpus, showing the mean, median, and mode lengths of adjectives and nouns by syllable count. These data provide an idea of the typical length of these words. Distributions of how often LENGTH is violated in the corpus data are in Table 3.16.

MEASURE	TOKEN	TYPE
Mean syllable count	3.3	3.7
Median syllable count	3	4
Mode syllable count	3 (31%)	4 (35%)

Table 3.14: Mean, median, and mode syllable counts for adjectives in the Italian Common Voice corpus.

MEASURE	TOKEN	TYPE
Mean syllable count	3.0	3.4
Median syllable count	3	3
Mode syllable count	3 (36%)	3 (35%)

Table 3.15: Mean, median, and mode syllable counts for nouns in the Italian Common Voice corpus.

LENGTH VIOLATION	FREQUENCY
Prenominal order	30,603 (42%)
Postnominal order	21,115 (29%)
Words are the same length	21,123 (29%)

Table 3.16: Violations of LENGTH among noun-adjective pairs (n=72841) in the Italian Common Voice corpus.

3.2.5 Consonant cluster constraints

The remaining phonological constraints analyzed in this work, VOICE and OCP-PLACE are not possible at word boundaries in Italian. Italian does not allow word-final consonants, so the voice and place features of adjacent consonants cannot be observed.

Constraint	Active status
Clash	Active across word boundaries (*retraction or doubling*; (Nespor and Vogel, 1979))
Lapse	Active across word boundaries (*beat addition*; (Nespor and Vogel, 1989))
Hiatus	Not active
Voice	Not possible
OCP-Place	Not possible
Length	Active for larger constituents (*object-verb-subject order*; (Cardinaletti, 2010))

Table 3.17: Summary table of which phonological constraints are active in Italian.

3.3 Methods

This thesis analyzes spoken corpus data from the Common Voice corpus, provided by Mozilla[3]. The Italian data analyzed in this work come from version it_317h_2021-07-21, consisting of 288 hours of validated speech from 6,407 speakers. Dialect information of the speakers was not reported; a native speaker listened to a random sample and concluded that there is likely a wide range of voices from various geographic areas. Noun-adjective pairs were extracted after the sentences were tagged using `spaCy`, whose models have a 97% accuracy on part-of-speech tagging for Italian[4].

The majority of the analysis was carried out on the phonemic representations of the audio in Common Voice. Common Voice is transcribed orthographically, and was converted to the phonemic level using lexical databases. The lexical database for Italian comes from PhonItalia (Goslin et al., 2014), and consists

[3]Accessed Fall 2021 at the following address: voice.mozilla.org

[4]Models were trained and tested on data from the Universal Dependencies Italian corpus (Bosco et al., 2013), the named-entity recognition Wikipedia corpus (Nothman et al., 2017), and lemma-token pairs, located at: github.com/michmech/lemmatization-lists In general, `spaCy` model accuracy was likely evaluated on held out data from Wikipedia, text, and formal speech.

of 120,000 word forms, created and manually checked by experts. 15% of the data were excluded due to missing pronunciations of one or both members of the noun-adjective pair (13239/88260)[5]. The phonological information in this database includes phonemes, syllable boundaries, and stress, all of which were used to code the constraints analyzed here. Specific constraint definitions are described in Table 3.18.

Constraint	Definition
Clash	Two stressed syllables adjacent at the word boundary.
Lapse	Three or more unstressed syllables across a word boundary.
Hiatus	Two members of the vowel set adjacent at the word boundary: {i e ɛ a o ɔ u}
Voice	Not possible for Italian.
OCP-Place	Not possible for Italian.

Table 3.18: Definition of phonological constraints for Italian.

The corpus data were split into two groups based on semantic similarity between the embeddings of the adjective in its prenominal and postnominal positions. The distribution of cosine similarities, shown in Figure 3.1, was fit for two distributions using a Gaussian mixture model. The boundary between these two distributions, marked with a red vertical line, was used to bin the data. Pair tokens with an adjective with a cosine similarity below 0.51 were categorized in the *similar* dataset, and those with a similarity above 0.51 were categorized in the *dissimilar* dataset.

A sample of 50 tokens in which clash is tolerated at the phonemic level between a noun and a flexible adjective were examined in Praat. These tokens were chosen out of a set of 97 pairs, which contained all pairs appearing in both or-

[5]The exclusion of this much data is not ideal, and while a more complete dataset may affect results, the overall analysis is expected to remain the same as there were no known patterns characterizing datapoints that were excluded.

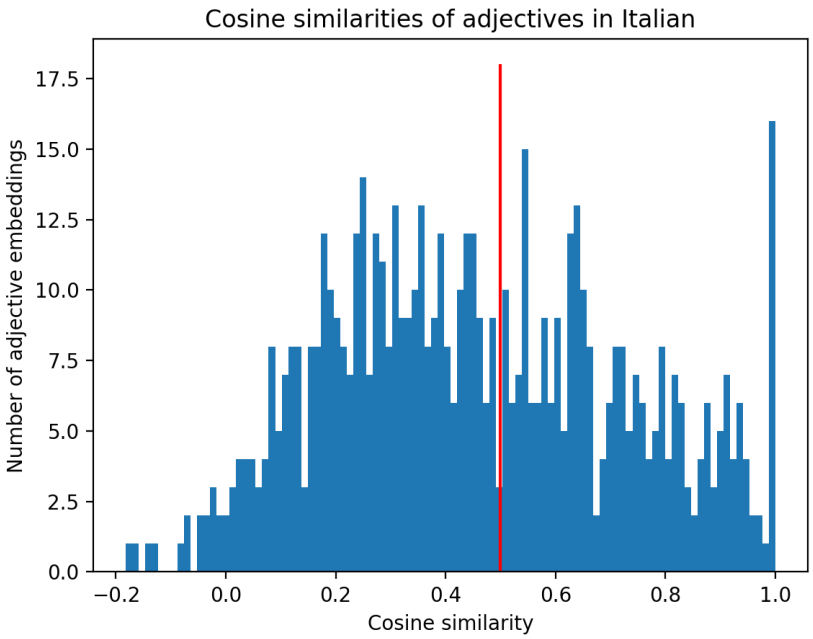


Figure 3.1: Distribution of cosine similarities measured between prenominal and postnominal embeddings of adjectives in Italian. Cutoff of 0.51 is marked with a red vertical line.

ders, as well as those with a flexible adjective that had a *similar* meaning in both positions, according to the methodology outlined above. Pairs were eliminated based on: clarity of the recording, perceived nativeness of the speaker, and bias for a diversity of pair types. A native speaker and myself analyzed the 50 recordings in Praat (Boersma and Weenik, 2022), to judge if any phonological repairs for clash in Italian were present in the tokens, namely stress shift or gemination of the onset of the second word in the pair (RS; see section 3.2.1).

3.4 Results

3.4.1 Regression models

A mixed effects logistic regression was fit to the Italian corpus data using `glmer` in R (R Core Team, 2016). The model predicted the order of pair tokens (prenominal ADJECTIVE NOUN or postnominal NOUN ADJECTIVE), using the phonological constraints possible at the word boundary in Italian: CLASH, LAPSE, and HIATUS, and LENGTH, in addition to RELATIVE FREQUENCY. For more information about how phonological constraints were coded, see 1.4.2. The model included random effects for ADJECTIVE and NOUN lemma identity[6]. See Figure 3.2 for a visualization of the results from both models at the end of this section.

$$\begin{aligned}&\texttt{glmer}(\text{outcome} \sim \textsc{Clash} + \textsc{Lapse} + \textsc{Hiatus} + \textsc{Length} + \textsc{Relative frequency} + \\ &\quad (1|\,\text{ADJ}) + (1|\,\text{NOUN}), \\ &\quad family = \text{“binomial”}\ )\end{aligned}$$

Regression models were fit separately for the two datasets separated by the semantic difference threshold, with the prediction that the phonological effects would be relatively stronger, or present only in the *similar* dataset which contains adjectives with a higher cosine similarity between their prenominal and postnominal embeddings. Results from the *dissimilar* model are presented in Table 3.19, and those from the *similar* model in Table 3.20.

In the *dissimilar* dataset, CLASH, LAPSE, HIATUS, LENGTH, and RELATIVE FREQUENCY are significant predictors of order. CLASH (-1) has a positive coefficient indicating that a violation of the constraint in prenominal order, but not in post-

[6]Random slopes were not included because the increase in the complexity caused the model to take too long to fit.

	Estimate	Std. Error	z value	p value
Intercept	**-1.38285**	**0.21641**	**-6.390**	**1.66e-10 *****
Constraint: Clash (-1)	**1.04242**	**0.37965**	**2.746**	**0.00604 ****
Constraint: Clash (1)	0.07387	0.17385	0.425	0.67090
Constraint: Lapse (-1)	0.10332	0.07099	1.456	0.14553
Constraint: Lapse (1)	**-0.40060**	**0.08051**	**-4.976**	**6.50e-07 *****
Constraint: Hiatus (-1)	**0.93908**	**0.09541**	**9.842**	**< 2e-16 *****
Constraint: Hiatus (1)	**-0.59527**	**0.10925**	**-5.449**	**5.07e-08 *****
Constraint: Length (-1)	**-0.35047**	**0.06600**	**-5.310**	**1.09e-07 *****
Constraint: Length (1)	**0.47741**	**0.06465**	**7.385**	**1.52e-13 *****
Relative frequency	**0.96145**	**0.17922**	**5.365**	**8.11e-08 *****

Table 3.19: Model fit for Italian data containing *less similar* adjectives. Number of observations is 39,568 noun-adjective pairs.

nominal order, actually correlates with prenominal order. Similarly, Lapse (1) has a negative coefficient, indicating that a violation in postnominal but not prenominal order correlates with postnominal order. The positive coefficient for Hiatus (-1) and the negative coefficient for Hiatus (1) also indicate tolerance of violations of the constraint where the phonologically-marked sequence could have been avoided. The results for Length do indicate avoidance of markedness. Length (-1) has a negative coefficient, indicating that noun-adjective pair with a violation of the constraint (e.g., long before short) in prenominal order, but not in postnominal order, is likelier to be postnominal compared to when the constraint is inactive. Showing a complementary effect, Length (1) has a positive coefficient, indicating that a pair with a violation of the constraint in postnominal order but not prenominal order is likelier to be prenominal compared to when the constraint is inactive. Relative frequency has a positive coefficient indicating that pairs with a greater degree of flexibility are prenominal-leaning, meaning the likelihood of a pair surfacing as prenominal increases as its flexibility increases.

	Estimate	Std. Error	z value	p value
Intercept	**-1.38285**	**0.21641**	**-6.390**	**1.66e-10 *****
Constraint: Clash (-1)	-1.2413	0.7343	-1.690	0.090952
Constraint: Clash (1)	0.5150	0.5085	1.013	0.311239
Constraint: Lapse (-1)	0.2076	0.1675	1.239	0.215235
Constraint: Lapse (1)	0.1306	0.1940	0.673	0.500851
Constraint: Hiatus (-1)	**0.7668**	**0.2233**	**3.434**	**0.000594 *****
Constraint: Hiatus (1)	**-0.8367**	**0.2425**	**-3.450**	**0.000561 *****
Constraint: Length (-1)	-0.2403	0.1639	-1.466	0.142551
Constraint: Length (1)	**0.6487**	**0.1589**	**4.081**	**4.48e-05 *****
Relative frequency	**-2.3063**	**0.6406**	**-3.600**	**0.000318 *****

Table 3.20: Model fit for Italian data containing *more similar* adjectives. Number of observations is 9,606 noun-adjective pairs.

In the *similar* dataset, Hiatus, Length, and Relative frequency are significant predictors of order. Like the *dissimilar* model, Hiatus (-1) has a positive coefficient and Hiatus (1) a negative coefficient, indicating a likelihood for tolerance of vowel-vowel sequences, rather than avoidance. Length (1) is significantly positive, indicating that a violation of the constraint in postnominal but not prenominal order correlates with prenominal order. Relative frequency has a negative coefficient, indicating that pairs with a greater degree of flexibility are postnominal-leaning, meaning the likelihood of a pair surfacing as postnominal increases as its flexibility increases.

A one-tailed Z-test was run to test the hypothesis that fixed effects coefficients are greater in the semantically similar model than in the dissimilar model, where coefficients are significant in both models (for more details, see section 1.4.3). The results of this test for Length (1) are in Table 3.21. The Z-test was not significant, indicating that the effect in the semantically similar model is not statistically greater than the effect in the dissimilar model.

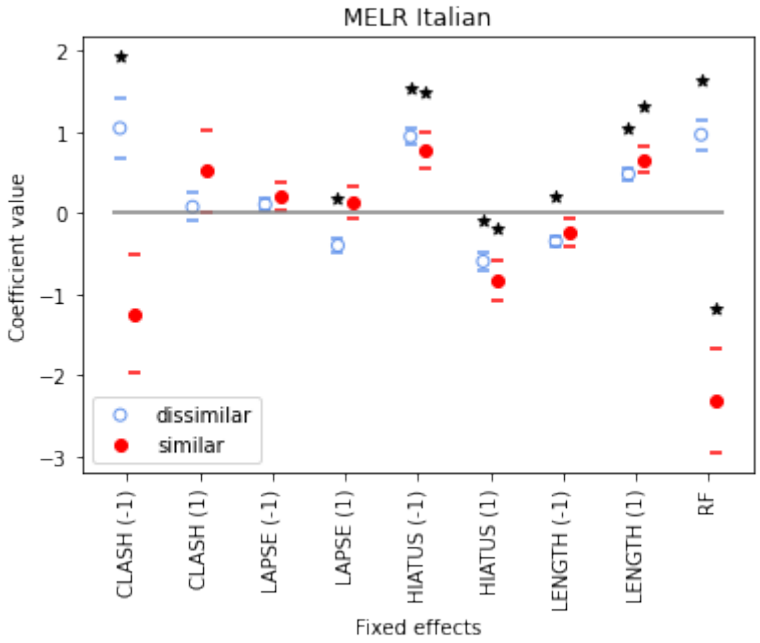


Figure 3.2: Visualization of the mixed-effects logistic regression results in Italian of the *dissimlar* (blue, triangle) and *similar* (red, circle) models. Coefficient values with standard error are shown for each fixed effect; significant effects are indicated by a star.

Fixed effect	Sim. model Estimate	Diss. model Estimate	Difference	p value
Length (1)	0.6487	0.47741	0.17129	0.159

Table 3.21: Results of a one-tailed Z-test for Italian that indicate whether or not the coefficient in the similar model is greater than the same coefficient in the dissimilar model.

3.4.2 Acoustic sample

50 tokens (41 types) of pairs with flexible adjectives that have underlying clash at the word boundary were examined at the phonetic level. A native speaker and myself agreed that 19/50 of the tokens had some type of clash repair, while the remaining 31 tokens did not have any repair – clash was produced in the token. Among the repairs, the plurality of them were cases of stress shift (8/19),

whereby final stress in the first word was less prominent before a stress-initial second word. This was due to a higher pitch, longer duration, or greater intensity on an underlyingly unstressed syllable earlier in word 1. See an example in Figure 3.3 of stress shift in the noun-adjective pair *città ricca*, 'rich city.' In this example, the final stress in *città* /t͡ʃi.ˈta/ is shifted somewhat to the first syllable, before *ricca* /ˈrik.ka/, which has initial stress that generates an underlying clash at the word boundary.

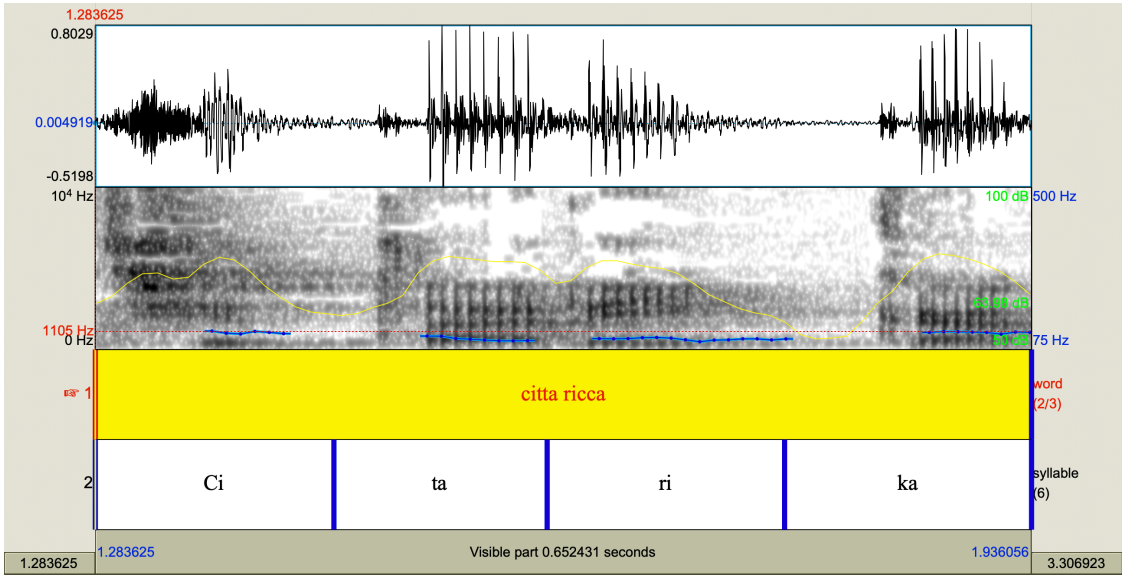


Figure 3.3: Stress shift in the noun-adjective phrase *città ricca* 'rich city', spoken by a male Italian speaker in the Common Voice corpus (file common_voice_it_19870780.mp3). Pitch is tracked in blue and intensity in yellow in the spectrogram.

There were also many cases of RS (7/19), meaning the onset of word 2 was lengthened, theorized as a repair that creates greater distance between two prominent syllables (Nespor and Vogel, 1979). This was a bit surprising given that RS does not happen after /r/ or in /sC/ onsets, meaning that only 32 out of the 50 tokens in the sample were possible candidates for RS; furthermore, RS is a regional phenomenon so it is possible that not all speakers in the sample have

RS in their grammar (Nespor and Vogel, 1979). An example of RS is shown in Figure 3.4, *città sola*. The duration of the /s/ in *sola* /ˈso.la/ is long in this clash environment.

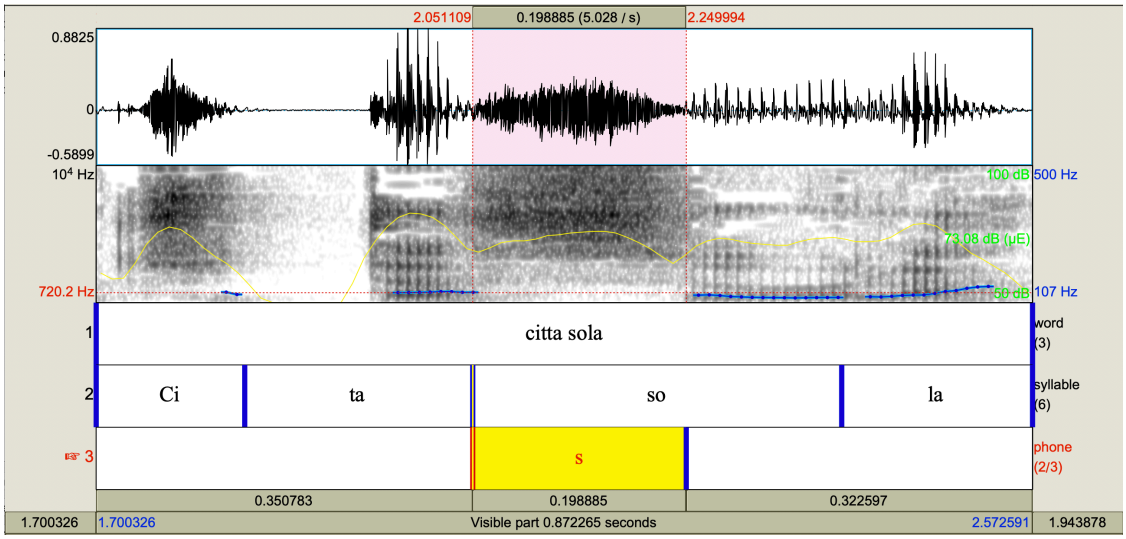


Figure 3.4: Onset lengthening in the phrase *città sola* 'lonely city', spoken by a male Italian speaker in the Common Voice corpus (file common_voice_it_25962916.mp3). Pitch is tracked in blue and intensity in yellow in the spectrogram.

Finally, there were two instances where the final vowel in adjective *migliore* before a noun with initial stress was realized. Normally, in prenominal position, *migliore* /miʎˈʎo.re/ is shortened to *miglior* /miʎˈʎor/, due to a process common in Standard Italian called *trocamento* (Meinschaefer, 2005). In a couple of cases of clash, however, this final vowel was produced, adding an additional stressless syllable where there would otherwise be a clash. See the example in Figure 3.5 of *miglior(e) film*.

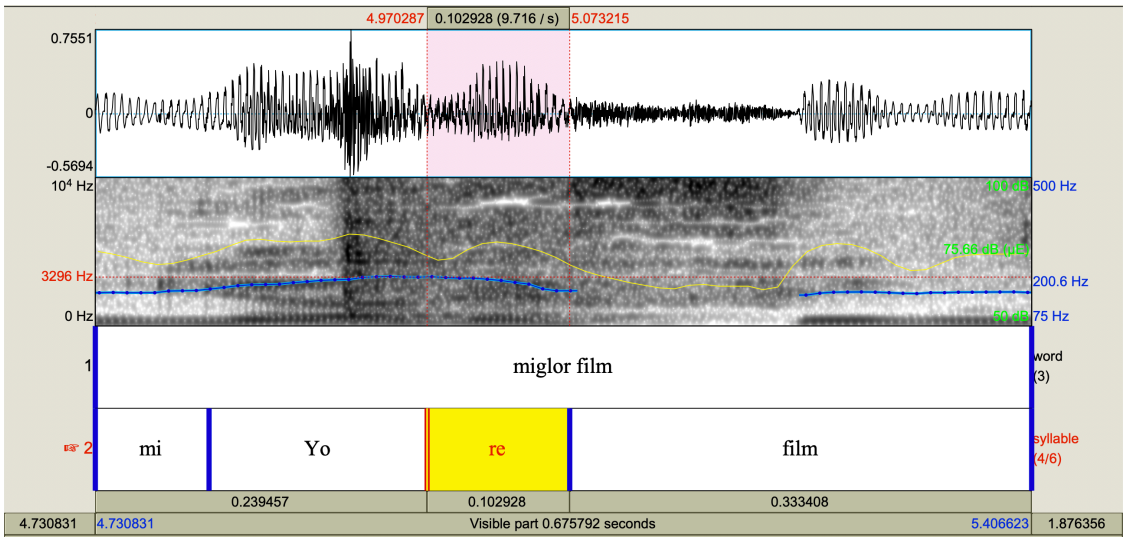


Figure 3.5: Final syllable of *miglior*(*e*) produced in the phrase *migliore film* 'best film', spoken by a female Italian speaker in the Common Voice corpus (file common_voice_it_20306010.mp3). Pitch is tracked in blue and intensity in yellow in the spectrogram.

Table 3.22 provides a summary of the results found in the acoustic sample of tolerated CLASH in Italian. The remaining two repairs were short pauses between words with clashes.

SURFACE PHENOMENON	FREQUENCY
Stress shift	8 (16%)
RS	7 (14%)
Other	4 (8%)
None	31 (62%)
TOTAL	50

Table 3.22: Results summary for the acoustic analysis of the sample of tolerated clashes in Italian.

3.5 Discussion

3.5.1 Regression models

Following the Phonologically-conditioned Hypothesis, the constraints Clash, Lapse, and Length were predicted to have significantly positive effects on noun-adjective word ordering, where the adjective is flexible. The evidence presented in section 3.2 supports the claim that these three constraints are active in Italian, at least at the word level, if not also between words; therefore, they were predicted to have an effect on noun-adjective ordering. These three effects were predicted to be significantly positive in at least the *similar* regression model, or to have a larger coefficient in the *similar* regression model compared to the *dissimilar* model, if significant in both, following the Semantically-conditioned Hypothesis.

These phonological predictions were true for Length. Both Length coefficients are significant in the dissimilar model, and Length (1) in the similar model. These results indicate that long before short sequences are likely to be avoided where possible.

Somewhat unexpectedly, Clash and Lapse have adverse effects on word order in the *dissimilar* model. Both of these phonologically-marked structures were expected to be significantly avoided via word ordering, given the evidence that they are otherwise phonologically avoided in Italian (Nespor and Vogel, 1979, 1989). A closer look at how often noun-adjective pair types are prenominal in the flexible noun-adjective pair dataset versus the dataset where clash is possible only in postnominal order (and thus prenominal order would constitute

an avoidance of clash via ordering) did not reveal that many of the pairs had a higher rate of prenominal order in the clash environment.

A key difference between these stress constraints constraints and LENGTH or HIATUS in French, however, is that it is possible to avoid instances of CLASH and LAPSE between nouns and adjectives using phonological repairs. Stress retraction or RS are possible repairs of clash at the word boundary, as is the insertion of an additional stress to repair lapse (Nespor and Vogel, 1979, 1989). It is not possible to insert extra syllables to repair a long before short sequence, nor is it possible in French, for instance, to delete a vowel or insert an epenthetic consonant to repair vowel hiatus[7]. LENGTH, and HIATUS in French, can be avoided only with word-order manipulation in the case of noun-adjective pairs. The acoustic sample was analyzed for clash repairs to test this hypothesis.

As predicted, word order is not used to avoid HIATUS; however, the coefficients are significantly correlated with tolerance of this constraint in both models. This appeared "preference" or tolerance for hiatus was not expected, but can be attributed to the combination of phonotactics and default word order of noun and adjective in the language[8]. Recall Tables 3.11 and 3.12, which report that 99% of adjectives and nouns in the Italian corpus end in a vowel; and only 18% of adjectives and 21% of nouns begin with a vowel in the full dataset (token frequency). This is in stark contrast to the phonological shape of words where hiatus is tolerated: 71% of adjectives (9831/13855) and 29% of nouns (4025/13855) are vowel-initial (token frequency), while the proportion of vowel-final words in both categories is still high at 99%. There is an increase in the proportion of

[7]This pertains to cases where liaison is not possible. For further details, results, and discussion of hiatus in French, please see Chapter 2.

[8]Models were re-run with a hiatus constraint that did not penalize high vowel-vowel sequences, in the event that these surface without hiatus as glide-vowel sequences. There was no appreciable difference in the coefficients between the models with this adjusted constraint.

vowel-initial words in both categories, but in adjectives in particular. Regarding word ordering, where hiatus is possible (including both tolerated and avoided occurrences), the majority of pairs are in postnominal NOUN ADJECTIVE order at 70% (13215/18834; token frequency). This is at an even higher frequency than in the overall dataset, which is 60% postnominal (43927/72841; token frequency). The regression result which shows a "preference" for hiatus – meaning where hiatus is possible, it is not avoided – is largely due to vowel-initial adjectives following nouns. It can be said that the default order in Italian is the hiatus-preferring order, and this is the driving force behind the significant result in the model which shows a preference for hiatus.

The semantic prediction that constraints be significantly positive or have a larger coefficient in the similar regression model compared to the dissimilar model, was not found in the data. LENGTH (-1) was significant only in the dissimilar model, and the results of the Z-test indicated that the effects of LENGTH (1) were not significantly greater in the similar model.

RELATIVE FREQUENCY is a significant predictor in both models; however, this coefficient is positive in the *dissimilar* model and negative in the *similar* model. As the semantic difference between an adjective in prenominal position versus postnominal position is minimized, flexible noun-adjective pairs are likelier to be postnominal, which is the default order. Inversely, if there is a greater semantic difference between positions of an adjective, flexible pairs are likelier to be prenominal, which is the position to which specific or special meanings of an adjective are usually attributed (Hall, 1948; Cinque, 2010). These facts also contribute to the larger negative coefficient for hiatus in the *similar* model, as this portion of the dataset is likelier to be postnominal, an order that prefers hiatus

as discussed above.

3.5.2 Acoustic sample

As discussed in the previous section, one hypothesis for the non-avoidance effect of Clash in Italian is that phonological repairs are preferred to different linear orderings, where possible. This does not appear to be the case: only 38% of the sample of tolerated phonological clashes had some sort of clash repair in the acoustic signal. Most of the time, speakers produced two prominent syllables across the noun-adjective word boundary, meaning neither word ordering nor phonological repair was utilized to avoid the surfacing of this phonologically-marked sequence.

An alternative proposal, then is that Clash is not as active in the phonology, and that the tendency of stress retraction described by Nespor and Vogel (1979) is not as strong as previously thought. If faithfulness to underlying lexical stress usually outranks stress clash, then we would not expect word ordering to be sensitive to this constraint, following the Phonologically-conditioned Hypothesis. Before drawing conclusions, however, I want to point out that the nature of looking at the acoustics of speech from a corpus varies greatly than that of looking at speech under a controlled, experimental environment. In an ideal world, extensive naturalistic corpus work would be evaluated in conjunction with thorough experimental testing to get a more complete understanding of speakers' phonology (see Cohn and Renwick, 2021, for an overview). Some common problems that arise in corpus work are present here. The tokens in the sample analyzed in this dissertation are produced by all different speakers with various dialects,

making the evaluation of empirical acoustic measurements such as duration, intensity, and f0 difficult because they cannot be compared to the production of the same token under a different environment within speakers. A recent experimental study of the acoustic effects of clash in Italian between nouns and color adjectives found that the final syllable in the first word had an increased duration and shifted vowel formants (higher F2) in a clash environment, compared to a non-clash environment (Burroni and Tilsen, 2022). It is possible that the less empirical, more impressionistic evaluation of clash in this dissertation obscured effects of stress shift revealed in experimental work. In this case, Clash may be optionally avoided via word ordering, given that it has been found to be phonologically avoided in a controlled environment.

CHAPTER 4

POLISH, HINDI, AND ARABIC

In order to begin building a typology of the nature of phonological markedness effects on noun-adjective ordering, this chapter presents analyses of three additional languages that differ in many respects from French and Italian. These languages are all outside of the Romance subgroup: Polish is a Slavic language (Indo-European); Hindi, an Indo-Aryan language (Indo-European); and Arabic, a Semitic language (Afro-Asiatic). In Polish and Hindi, adjectives typically precede the noun they modify; postnominal is the default in Arabic, like French and Italian. These additional languages also have important phonological differences: complex syllable structure is observed in Polish, and to some extent in Hindi; and, Arabic and Hindi both have weight-sensitive stress systems. Syntactically, Polish, Hindi, and Arabic all exhibit much freer word order than French and Italian. While Polish, French, and Italian are all subject-verb-object languages, Arabic and Hindi differ in their basic word orders: Arabic is a VSO language[1] and Hindi SOV.

This chapter primarily discusses aspects of Polish as it is spoken by the educated population in Poland; the standardized form of Hindi, also known as Modern Standard Hindi (Kachru, 2006); and, Modern Standard Arabic. Arabic as an object of study is particularly challenging. There are significant linguistic differences between so-called "dialects" of Arabic, which in turn differ greatly from the standard. Additionally, Modern Standard Arabic is considered to be a learned language or a *lingua franca* among speakers of vernacular Arabic varieties, and so does not have a true speech community (Kamusella, 2017).

[1]Dialectal varities of Arabic have been described as SVO (Dryer and Haspelmath, 2013).

The choice of Modern Standard Arabic for this analysis was governed by the resources available: a desire to use the same corpus across all five languages. Future work should replicate this study using data from a specific spoken dialect of Arabic for a better understanding of the speaker's grammar.

The additional analyses of these languages offer greater breadth, rather than depth, to this dissertation and are more exploratory. There are generally fewer resources available for these three languages, especially compared to French, in terms of speech corpus data, phonological dictionaries, and part-of-speech tagging. This rest of this chapter is organized as follows. Noun-adjective ordering and its implications are discussed in the next section. Section 4.2 details how the phonological constraints are treated in the phonology of each language, including predictions about the corpus data. Section 4.3 provides language-specific methodological details. Results are presented and discussed in section 4.4, language by language.

4.1 Noun-adjective flexibility

In Polish, the canonical ordering of adjectives with respect to the noun that they modify is prenominal: ADJECTIVE NOUN, but adjectives can also appear in postnominal position. 68% of noun-adjective pairs with a flexible adjective appear in prenominal order (6499/9601), confirming this description. Prenominal position has been associated with descriptive, qualitative adjectives and postnominal with classifying adjectives (Sadowska, 2012; Swan, 2002). The same adjective can appear before or after the noun, and may involve a difference in meaning between describing and classifying the noun. See the examples in (24), from

Swan (2002), p.127.

(24) **Flexible adjectives in Polish**

a. zwykły kleszcz ∼ kleszcz zwykły
ordinary tick tick ordinary
'an ordinary tick' ∼ 'The Common Tick'

b. teatralny gest ∼ gest teatralny
theatrical gesture gesture theatrical
'a theatrical gesture' ∼ 'a gesture used in the theater'

c. piękna literatura ∼ literatura piękna
beautiful literature literature beautiful
'beautiful literature' ∼ '*belles lettres*'

In Hindi, the canonical ordering of adjectives with respect to the noun that they modify is also prenominal: ADJECTIVE NOUN. Corpus results confirm this, with 70% of noun-adjective pairs containing a flexible adjective appearing in prenominal order (169/242). Flexible word order in Hindi allows for the majority of adjectives to also appear in postnominal position (Jain, 1995; Kachru, 2006). Prenominal position is generally associated with attributive adjectives, and postnominal with non-attributive adjectives or predicates (Jain, 1995; Kachru, 2006). Both orders are exemplified in (25)[2].

(25) **Flexible adjectives in Hindi**

a. Maine laal kitaab khareedi
I red book bought
'I bought a red book.'

b. Maine kitaab laal khareedi
I book red bought
'I bought a *red* book (not any other color).'

[2]Thanks to Bhavya Pant for these examples, as well as their transliterations from Devanagari.

Adjectives typically follow nouns in Arabic (Ryding, 2005). This is reflected in the corpus data: 70% of noun-adjective pairs with a flexible adjective are in postnominal order (5318/7606). Prenominal order is also possible, but has been reported to invoke a different meaning in the adjective or phrase. As can be seen in the examples in (26), the adjective in postnominal position has an attributive meaning, but a non-attributive one in prenominal position (Kremers, 2003, p.59 reproduced from El-Ayoubi et al., 2001).

(26) **Flexible adjectives in Arabic**

a. maᶜa ğazīl-i -l-šukr-i
with abundant-GEN the-thanks-GEN
'with the greatest (of) thanks'

a′. šukr-an ğazīl-an
thanks-ACC abundant-ACC
'many thanks'

b. sābiq-u ʾinḏār-i-n
preceding-NOM warning-GEN-INDEF
'a fore-warning (lit. 'the preceding of a warning')

b′. ʾinḏār-u-n sābiq-u-n
warning-NOM-INDEF preceding-NOM-INDEF
'a previous warning (lit. 'a preceding warning')

4.2 Phonology

In this section, I discuss the phonological markedness constraints as they pertain to Polish, Hindi, and Modern Standard Arabic: whether or not they can be violated at the word boundary, and if so whether they are repaired phonologically. A summary of the constraints is provided at the end of this section, in Table 4.1.

The following sources are the basis of the general understanding of the phonology of each language: for Polish, Sadowska (2012); Swan (2002); Hindi, Kachru (2006); Ohala (1983); and Arabic, Ryding (2005).

4.2.1 Stress clash

Hayes and Puppel (2019) detail the rhythm rule in Polish, but at the level of the foot rather than the syllable. Two directly adjacent stressed syllables do not occur in the language, since Polish does not have final stress (though it has initial). The rhythm rule applies then, to two adjacent feet with primary stress. This is not at the same prosodic level as the clash constraint as I define it in this dissertation, which is at the level of the syllable; therefore, clash is not a possible predictor of word ordering in Polish.

Pandey (2021) reports that avoidance of clash word-internally is variable in Hindi. Clash tolerance may differ by dialect as well: in her Optimality Theoretic analysis of the stress system of Hindi, Buchanan (2012) argues that *Clash is ranked higher in the stress system of the Eastern Standard dialect compared to Kelkar's Hindi. Given these facts, it is difficult to predict whether clash will have an effect on word ordering in Hindi.

Clash is possible in Arabic; however, word-final stress is relatively rare. Stress is assigned to the final syllable in the event that it is superheavy, which is defined as a syllable containing a long vowel and a coda, or a complex coda. Only 528 of the 7,606 noun-adjective pairs with a flexible adjective contain a word with final stress (7%). There are no previous reports in the literature of clash avoidance in Arabic, therefore it is not predicted to have an effect on word ordering.

4.2.2 Stress lapse

Like clash, lapses in Polish are also defined over feet and their dispreference is not straightforward (Newlin-Łukowicz, 2012). LAPSE is also considered not to be a possible predictor of word ordering in this language.

Previous literature does not report a dispreference for lapse in Hindi, though it is possible within words and across the word boundary. For this reason, it is not predicted to have a significant effect on word ordering.

In Palestinian Arabic, three unstressed syllables in word final position trigger a stress shift, an effect of lapse (Houghton, 2008); however, no such shift has been reported for Modern Standard Arabic, so no effect on word ordering is predicted.

4.2.3 Vowel hiatus

Vowel hiatus is repaired by glottal stop in Polish, as found in an experimental study by Schwartz (2013). In a production experiment on phrase-internal vowel-vowel sequences across a word boundary, 76% were produced with glottalization, indicating that this repair is fairly robust. Vowel-initial words in Polish can also be produced with glottalization following a consonant, but less often than when preceded by a vowel (Malisz et al., 2013; Schwartz, 2013). Hiatus is therefore predicted to condition noun-adjective ordering in Polish.

Sequences of vowels are repaired with the insertion of glides /j/, /w/, or /ʊ/ in Hindi (Singh and Sarma, 2011; Kachru, 2006; Ohala, 1983). This is true for stem-internal instances of hiatus, such as underlying /kɔ.ɑ/ pronounced [kəu.vɑ]'crow'; as well as hiatus derived via verbal morphology, such as /kʰa-

/a/ pronounced [kʰa.ja] 'eat-PERFECT' (Kachru, 2006, p.31). Hiatus is predicted to affect word ordering in Hindi, given the evidence of this process of epenthesis as a means to phonologically avoid hiatus.

Onsets are required word-initially in Arabic, so hiatus at the word boundary is not possible (Ryding, 2005). It is not included as a predictor in the Arabic model.

4.2.4 Voice-disagreeing clusters

Regressive voicing assimilation has been reported for obstruent clusters in Polish, both within and across words; however, clusters containing consonants that both belong to the same word show voicing assimilation much more reliably than those clusters that are across a word boundary (Gussmann, 1992). Polish also has word-final devoicing, but assimilation of consonants across the word boundary happens whether the coda is phonologically voiced or voiceless (Gussmann, 1992).

Ohala (1983) reports that consonant clusters that disagree in voicing are disallowed in Hindi. In a study of spontaneous speech, Ohala (2001) finds assimilation in voicing in consonant clusters across morpheme and word boundaries.

Regressive voicing assimilation is reported in many Arabic dialects (Egyptian, Sudanese, and Daragözü, Abu-Mansour, 1996; Palestinian, Tamim, 2017; Cairene Kabrah et al., 2011), as well as in Modern Standard Arabic (Altakhaineh and Zibin, 2014).

Phonological repairs of voice-disagreeing clusters in Polish, Hindi, and Ara-

bic leads to the prediction that VOICE will be a predictor of word ordering in all three languages, following the PHONOLOGICALLY-CONDITIONED HYPOTHESIS.

4.2.5 OCP-Place

Hindi has phonotactic restrictions against consonant clusters at the same place of articulation, as well as phonological processes that are sensitive to them. Ohala (1983) states that "initially, medially, and finally, two stops of the same point of articulation do not follow each other" (p.56). Additionally, schwa deletion is blocked between consonants at the same place of articulation: see /t^həpək+iː/ → [t^həpkiː] 'a pat' but /aːdət+ẽː/ → [aːdətẽː] 'habits' (Baković, 2005, p.300).

The co-occurrence restriction on consonants with the same place of articulation has been widely studied in Arabic (Greenberg, 1950; Pierrehumbert, 1993; McCarthy, 1994; Frisch and Zawaydeh, 2001). Virtually no Arabic roots contain adjacent identical consonants, and non-identical adjacent consonants with the same place of articulation are uncommon (Pierrehumbert, 1993). Faiq and Burhanuddin (2019) describe various dissimilation processes that target identical consonant sequences in Arabic, and McCarthy (1986) shows how metathesis is blocked when it would allow false geminates (identical, heteromorphemic consonants) to surface.

No effect of the constraint OCP-PLACE on noun-adjective ordering is predicted for Polish, a language in which no dispreference for clusters at the same place of articulation has been documented; effects are predicted for Hindi and Arabic. This is following the PHONOLOGICALLY-CONDITIONED HYPOTHESIS.

4.2.6 Length

Some effects of constituent length on word order have been reported for Polish (Siewierska, 1993). In a corpus study of the relative lengths of subject and object constituents by number of words, Siewierska found that short before long was more common than long before short for the most common sentence structure, SVO, as well as some other sentence structures possible in the language: SOV, VSO, and VOS. Given this evidence, length is expected to have an effect on word ordering in Polish.

It has been noted that prosodic end-weight effects are non-existent or even reversed in verb-final languages (i.e., languages with SOV or OSV dominant constituent order; Ryan, 2019). Hindi is the only verb-final language among those investigated in this dissertation, having a dominant SOV word order (McGregor, 1977); this constraint is therefore predicted to have no effect or a reverse effect on noun-adjective word ordering in this language.

There is evidence of heavy NP shift in Arabic. In a corpus study of Modern Standard Arabic, Mohamed (2014) finds that the chances of an object NP shift go up as its size increases (defined as number of morphemes) and vice versa, in parallel with the heavy NP shift phenomenon in English. In the same study, similar results were found for subject NP shift. An effect of length on noun-adjective ordering in Arabic is therefore expected.

Constraint	Active status
Clash	Not possible for Polish or Arabic, active in Hindi
Lapse	Not active in Hindi or Arabic, not possible in Polish
Hiatus	Active in Polish and Hindi, not possible in Arabic
Voice	Active in Polish, Hindi, and Arabic
OCP-Place	Active in Hindi and Arabic, but not Polish
Length	Active in Polish and Arabic, but not Hindi

Table 4.1: Summary table of which phonological constraints are active in Polish, Hindi, and Arabic.

4.3 Methods

4.3.1 Polish

The Polish data analyzed in this work come from version pl_152h_2021 of the Common Voice corpus, consisting of 129 hours of validated speech from 2,918 speakers. Dialect information of the speakers was not available. Noun-adjective pairs were extracted after the sentences were tagged using `spaCy`, whose models have a 98% accuracy on part-of-speech tagging for Polish[3].

The majority of the analysis was carried out on the phonemic representations of the audio in Common Voice. Common Voice is transcribed orthographically, and was converted to the phonemic level using lexical databases. The lexical database for Polish comes from WikiPron, and consists of 86,000 word forms

[3]Models were trained and tested on data from the Universal Dependencies Polish corpus (Wróblewska, 2018), the National Corpus of Polish (Przepiórkowski, 2012), and PoliMorf Woliński et al. (2012). In general, `spaCy` model accuracy was likely evaluated on held out data from text, dictionaries, and formal speech.

scraped from Wiktionary (Lee et al., 2020). 59% of the data had to be excluded due to missing pronunciations of one or both members of the noun-adjective pair (18102/30558 pairs)[4]. The phonological information in this database includes phonemes, syllable boundaries, and stress, all of which were used to code the constraints analyzed here. Specific constraint definitions are described in Table 4.2.

Constraint	Definition
Clash	Not defined for Polish.
Lapse	Not defined for Polish.
Hiatus	Two members of the vowel set adjacent at the word boundary: {i ɛ ɨ a u ɔ ɛ̃ ɔ̃}.
Voice	One voiceless consonant and one voiced consonant adjacent at the word boundary: {p t k kʲ t͡s t͡ʂ t͡ɕ f s ʂ ɕ x xʲ}, {b d g gʲ d͡z d͡ʐ t͡ʑ v z ʐ ʑ m n ɲ r l j w}
OCP-Place	Two consonants at the same place of articulation adjacent at the word boundary: {p b f v w}, {t t͡s d d͡z s z r l}, {t͡ʂ d͡ʐ ʂ ʐ} {t͡ɕ t͡ʑ ɕ ʑ j}, {k kʲ g gʲ x xʲ w}

Table 4.2: Definition of phonological constraints for Polish.

4.3.2 Hindi

The Hindi data analyzed in this work come from version hi_11h_2021-07-21 of the Common Voice corpus, consisting of 8 hours of validated speech from 214 speakers. Dialect information of the speakers was not available. Noun-adjective pairs were extracted after the sentences were tagged using `Stanza`, a part-of-speech tagger trained on the Universal Dependencies corpus of Hindi, with a reported accuracy of 98% (Qi et al., 2020; Bhat et al., 2017)[5].

[4]The exclusion of more than half of the data due to missing pronunciations is far from ideal, and these results should be interpreted with this methodological caveat in mind. A more complete analysis may result in different trends observed in the data.

[5]Accessed Fall 2021 at github.com/stanfordnlp/stanza.

The majority of the analysis was carried out on the phonemic representations of the audio in Common Voice. The lexical database for Hindi comes from WikiPron, and consists of over 13,000 word forms scraped from Wiktionary (Lee et al., 2020). 67% of the data had to be excluded due to missing pronunciations of one or both members of the noun-adjective pair (523/1596 pairs). The phonological information in this database includes phonemes and syllable boundaries which were used to code the constraints analyzed here. Stress was added automatically to the forms using the following rules from Kelkar (1968): (1) Stress is assigned to the heaviest syllable in a word, (2) In the case of a tie, stress is assigned to the right-most, non-final syllable of those that are tied. Specific constraint definitions are described in Table 4.3. The Hindi dataset had the greatest challenges of all five languages analyzed in this dissertation. The exclusion of more than half of the data as well as the automatic stress assignment to forms are a considerable methodological weakness in Hindi, and these results should be interpreted with this caveat in mind.

Constraint	Definition
Clash	Two stressed syllables adjacent at the word boundary.
Lapse	Three or more unstressed syllables across a word boundary.
Hiatus	Two members of the vowel set adjacent at the word boundary: {i ɪ e ɛ ə æ u ʊ o ɔ a}, which can also occur lengthened and/or nasalized.
Voice	One voiceless consonant and one voiced consonant adjacent at the word boundary: {p pʰ f t tʰ s ʈ ʈʰ ʂ t͡ʃ t͡ʃʰ ʃ k kʰ x q}, {b bʱ v ʊ d dʱ z r l ɖ ɖʱ ɽ ɽʱ ɭ d͡ʒ d͡ʒʱ ʒ j ɦ}
OCP-Place	Two consonants at the same place of articulation adjacent at the word boundary: {p pʰ b bʱ f v ʋ}, {t tʰ d dʱ s z r l}, {ʈ ʈʰ ɖ ɖʱ ʂ ɽ ɽʱ ɭ}, {t͡ʃ t͡ʃʰ d͡ʒ d͡ʒʱ ʃ ʒ j}, {k kʰ g gʱ x ɣ}, {q}, {ɦ}

Table 4.3: Definition of phonological constraints for Hindi.

There are considerably fewer datapoints in Hindi compared to the other lan-

guages analyzed in this dissertation. After exclusions due to missing pronunciations, and elimination of pairs without a flexible adjective, only 242 noun-adjective pair tokens remain. Because there were so few datapoints, the semantic analysis of the Hindi data was not performed, and the mixed effects model of this dataset is inconclusive.

4.3.3 Arabic

The Arabic data analyzed in this work come from version ar_137h_2021-07-21 of the Common Voice corpus, consisting of 85 hours of validated speech from 1,052 speakers. Dialect information of the speakers was not available, but a native speaker confirmed that a random sample of the data[6] was undoubtedly Modern Standard Arabic.

Sentences in the corpus are transcribed in the traditional Arabic script. This orthography was converted to Buckwalter representations and tagged for part-of-speech using MADAMIRA (Pasha et al., 2014). Since the majority of the analysis was carried out on the phonemic representations, the Buckwalter transliteration was converted to IPA using a script that I wrote alongside another linguist, who is a native speaker of Arabic (Buckwalter, 2004; Hassan Munshi, personal communication)[7]. Because of this rule-based method, no data had to be excluded due to missing pronunciations.

Syllabification and stress were assigned automatically, following the organization of syllables in Ryding (2005), and stress assignment described by Watson (2011): (1) Stress is assigned to the final superheavy syllable, (2) Stress is as-

[6]Random sample was approximately 50 datapoints.
[7]This script is publicly available at github.com/katherineblake/language-scripts.

signed to the heavy penultimate syllable, (3) Otherwise stress is assigned to the antepenultimate syllable (or initial if disyllabic). Specific constraint definitions are described in Table 4.4.

Constraint	Definition
Clash	Two stressed syllables adjacent at the word boundary.
Lapse	Three or more unstressed syllables across a word boundary.
Hiatus	Not possible for Arabic.
Voice	One voiceless consonant and one voiced consonant adjacent at the word boundary: {f t s tˤ sˤ θ ʃ k q χ ħ ʔ}, {m d z dˤ n r l zˤ ð g ʁ ʕ}
OCP-Place	Two consonants at the same place of articulation adjacent at the word boundary: {b m f}, {n r l}, {t d s z tˤ dˤ sˤ zˤ θ ð ʃ}, {k g q}, {χ ʁ ħ ʕ h ʔ}[8]

Table 4.4: Definition of phonological constraints for Arabic.

4.4 Results and discussion

Mixed effects logistic regression models were fit to the corpus data for each language using `glmer` in R (R Core Team, 2016). The models predicted the order of pair tokens (prenominal adjective noun or postnominal noun adjective), using the phonological constraints possible at the word boundaries in each language: Clash, Lapse, Hiatus, Voice, OCP, and/or Length, in addition to Relative Frequency. For more information about how phonological constraints were coded, see 1.4.2. The models also included random intercepts for Adjective and Noun

[8]Constraint definition based on that proposed by (Pierrehumbert, 1993) for consonant restrictions in Arabic roots.

lemmas[9].

$$\texttt{glmer}(\text{outcome} \sim \textsc{Clash} + \textsc{Lapse} + \textsc{Hiatus} + \textsc{Voice} + \text{OCP} + \textsc{Length} + \textsc{Relative frequency} + (1|\ \text{ADJ}) + (1|\ \text{NOUN}), \ family = \text{“binomial”}\ )$$

Results are presented and discussed for each language in turn in the remainder of this section.

4.4.1 Polish

In Polish, a regression model was fit to the dataset including all noun-adjective pairs with flexible adjectives; flexible being defined as occurring in both prenominal and postnominal position in the corpus. Individual models of the *similar* and *dissimilar* datasets had convergence issues, so the model of these combined datasets is reported in Table 4.5. A visualization of the results is provided in Figure 4.1.

In Polish, Voice, OCP, and Length are all significant predictors of the ordering of nouns and adjectives, with positive coefficients. For Voice, OCP, and Length (-1), this indicates that a violation of the constraint (e.g., consonants with a mismatch in voicing at the word boundary) in prenominal order, but not in postnominal order, correlates with postnominal order. A positive coefficient for Length(1) indicates that a violation in postnominal but not prenominal correlates with prenominal order. Relative frequency is also not significant, indicating that there is not a relationship between pair flexibility and prenominal

[9]Random slopes were not included because the increase in the complexity caused the model to take too long to fit.

	Estimate	Std. Error	z value	p value
Intercept	**7.63927**	**0.61491**	**12.423**	**< 2e-16 *****
Constraint: Hiatus (-1)	0.06261	0.40511	0.155	0.877179
Constraint: Hiatus (1)	-0.51850	0.33478	-1.549	0.121428
Constraint: Voice (-1)	**0.34304**	**0.16082**	**2.133**	**0.032924 ***
Constraint: Voice (1)	-0.05928	0.14670	-0.404	0.686158
Constraint: OCP (-1)	**-2.92285**	**0.55504**	**-5.266**	**1.39e-07 *****
Constraint: OCP (1)	0.23210	0.22824	1.017	0.309201
Constraint: Length (-1)	**-0.89658**	**0.13903**	**-6.449**	**1.13e-10 *****
Constraint: Length (1)	**-0.45729**	**0.19246**	**-2.376**	**0.017500 ***
Relative frequency	**1.59593**	**0.42694**	**3.738**	**0.000185 *****

Table 4.5: Model fit for Polish data containing flexible adjectives. Number of observations is 9,601 noun-adjective pairs.

versus postnominal ordering.

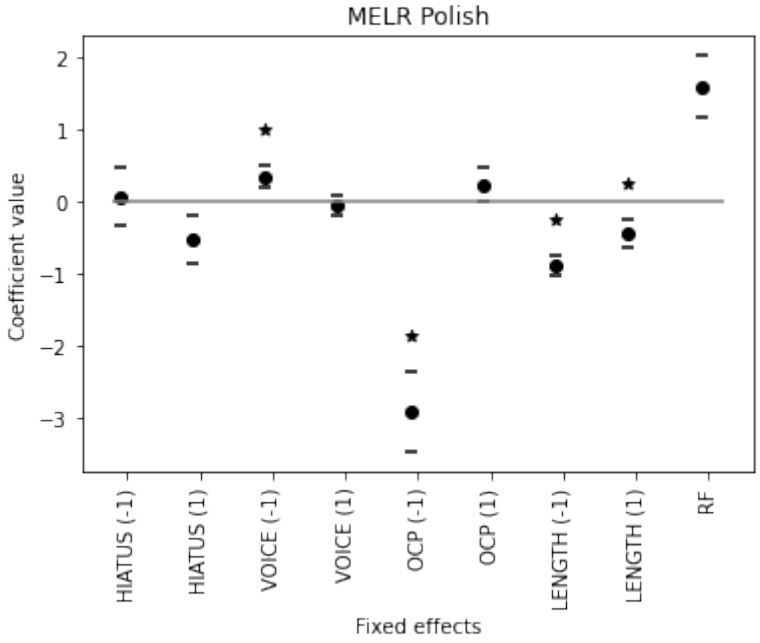


Figure 4.1: Visualization of the mixed-effects logistic regression results in Polish. Coefficient values with standard error are shown for each fixed effect; significant effects are indicated by a star.

Discussion

Following the PHONOLOGICALLY-CONDITIONED HYPOTHESIS, the constraints VOICE, HIATUS, and LENGTH were expected to have an effect on noun-adjective ordering in Polish, whereas no effect was predicted for OCP. Like French, results supported the hypothesis for VOICE (-1), but were inconclusive for VOICE (1) so it is unclear if there is a main effect of avoiding VOICE violations via word ordering. Like French, VOICE can be repaired phonologically at the word boundary via regressive voicing assimilation in Polish. A look at the acoustics of noun-adjective pairs where voiceless-voiced or voiceless-voiced sequences are not avoided, as was done in French, may reveal surface-level repairs. Though HIATUS was predicted to have an effect on word ordering, it is phonologically reparable at the word boundary with glottalization. Thus it follows the pattern of VOICE in French, and CLASH and LAPSE in Italian. Future work should include an acoustic analysis of noun-adjective pairs that violate HIATUS in Polish, as was done for similarly-behaving constraints in French and Italian, to investigate whether vowel-vowel sequences are produced with a phonological repair by speakers. Again, no effect was predicted for OCP and was found to be significant for only one of the simple effects, OCP (-1), meaning consonants at the same place of articulation are avoided when they occur in prenominal order, but not significantly in postnominal order. On the other hand, LENGTH was predicted to be avoided and while this was found in postnominal order, LENGTH (-1), the opposite was true for prenominal order, LENGTH (1). Future work should look at the occurrence and avoidance of OCP and LENGTH in Polish in greater detail.

4.4.2 Hindi

For Hindi, there was relatively little data for the analysis. There are only 242 noun-adjective pairs in the models presented here, compared to thousands or tens of thousands of pairs in the other languages analyzed in this dissertation. This is due to fewer hours of speech to begin with: 8 hours of Hindi are available through Common Voice compared to over 100 hours each for French, Italian, and Polish. Additionally, over half of the noun-adjective pairs extracted from Common Voice had to be eliminated due to lack of phonological forms for one or both of the words in the pair. Finally, the amount of data was halved again once noun-adjective pairs without a flexible adjective were filtered out. The 242 flexible adjective tokens are comprised of 33 types; this frequency distribution is shown in Figure 4.2. There is not a single adjective or group of adjectives dominating the tokens, suggesting that this is a small but potentially fairly representative sample of a larger dataset in Hindi that could be analyzed in the future.

I show the results of the mixed effects logistic regression model even though it did not converge in the hopes that, with more data, the same analysis can be carried out in the future with greater success. I will not further discuss the model output in Table 4.6, but it is included here for completeness.

Discussion

There was unfortunately not enough data to draw any conclusions about the use of word order manipulation to avoid phonologically-marked structures in Hindi noun-adjective pairs. Following the Phonologically-conditioned Hypothesis, Hiatus, Voice, and OCP were predicted to be avoided. Whether Clash is active

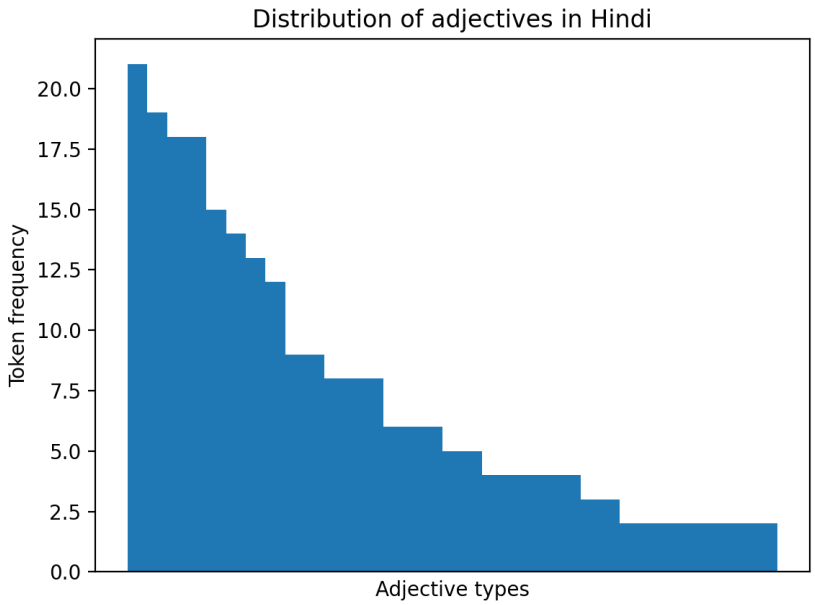


Figure 4.2: Token frequencies of adjective types in Hindi. 242 total tokens and 33 types.

	Estimate	Std. Error	z value	p value
Intercept	-3.05759	4.00757	-0.763	0.4455
Clash (-1)	0.21309	0.83869	0.254	0.7994
Clash (1)	0.31021	0.73559	0.422	0.6732
Lapse (-1)	1.40308	1.40240	1.000	0.3171
Lapse (1)	-18.56526	2747.84163	-0.007	0.9946
Hiatus (-1)	2.31464	2.04384	1.132	0.2574
Hiatus (1)	1.62307	1.43635	1.130	0.2585
Voice (-1)	**-2.02698**	**0.90620**	**-2.237**	**0.0253 ***
Voice (1)	**-1.62527**	**0.67275**	**-2.416**	**0.0157 ***
OCP (-1)	0.45935	1.50202	0.306	0.7597
OCP (1)	0.52853	1.10931	0.476	0.6338
Length (-1)	0.03220	0.75985	0.042	0.9662
Length (1)	-0.03709	0.91082	-0.041	0.9675
Relative frequency	4.73418	3.96662	1.194	0.2327

Table 4.6: Model fit for Hindi data containing flexible adjectives. Number of observations is 242 noun-adjective pairs. Model did not converge.

in the language's phonology is unclear (Pandey, 2021; Buchanan, 2012, see 4.2 for more details). Lapse was not predicted to have an effect.

Unlike the other languages analyzed in this dissertation, Length was predicted to have no effect, or to be significantly negative in Hindi. This is due to the constituent order of Hindi, which is verb-final (McGregor, 1977). In languages with this right-branching syntactic structure, end-weight effects have been found to be non-existent or reversed (Ryan, 2019). Future work using larger corpora may confirm or disprove these hypotheses.

4.4.3 Arabic

In Arabic, a regression model was fit to the dataset consisting of all noun-adjective pairs with flexible adjectives. Individual models of the *similar* and *dissimilar* datasets had convergence issues, so the model of these combined datasets is reported in Table 4.7. A visualization of the results is provided in Figure 4.3.

	Estimate	Std. Error	z value	p value
Intercept	**2.1036**	**0.8753**	**2.403**	**0.01625 ***
Constraint: Clash (-1)	-1.2588	0.8121	-1.550	0.12116
Constraint: Clash (1)	**2.2080**	**0.7972**	**2.770**	**0.00561 ****
Constraint: Lapse (-1)	**0.8630**	**0.1878**	**4.595**	**4.34e-06 *****
Constraint: Lapse (1)	-0.2181	0.1813	-1.203	0.22896
Constraint: Voice (-1)	**-6.2449**	**0.3828**	**-16.313**	**< 2e-16 *****
Constraint: Voice (1)	**0.6551**	**0.2142**	**3.059**	**0.00222 ****
Constraint: OCP (-1)	**-4.3437**	**0.5068**	**-8.570**	**< 2e-16 *****
Constraint: OCP (1)	**0.8339**	**0.4020**	**2.074**	**0.03807 ***
Constraint: Length (-1)	**-1.2309**	**0.1865**	**-6.601**	**4.08e-11 *****
Constraint: Length (1)	**1.6072**	**0.2251**	**7.139**	**9.43e-13 *****
Relative frequency	**-4.2333**	**0.8315**	**-5.091**	**3.56e-07 *****

Table 4.7: Model fit for Arabic data containing flexible adjectives. Number of observations is 7,606 noun-adjective pairs.

In Arabic, at least one simple effect for each fixed effect is a significant predictor of the ordering of nouns and adjectives. Clash, Voice, OCP, and Length have coefficients that align with the avoidance of violations of these constraints via word ordering. Lapse (-1), however, has a positive coefficient, meaning lapses across the word boundary are tolerated rather than avoided. Relative frequency is also negative, corresponding to a correlation between flexibility of a pair and a tendency to be in postnominal, default order.

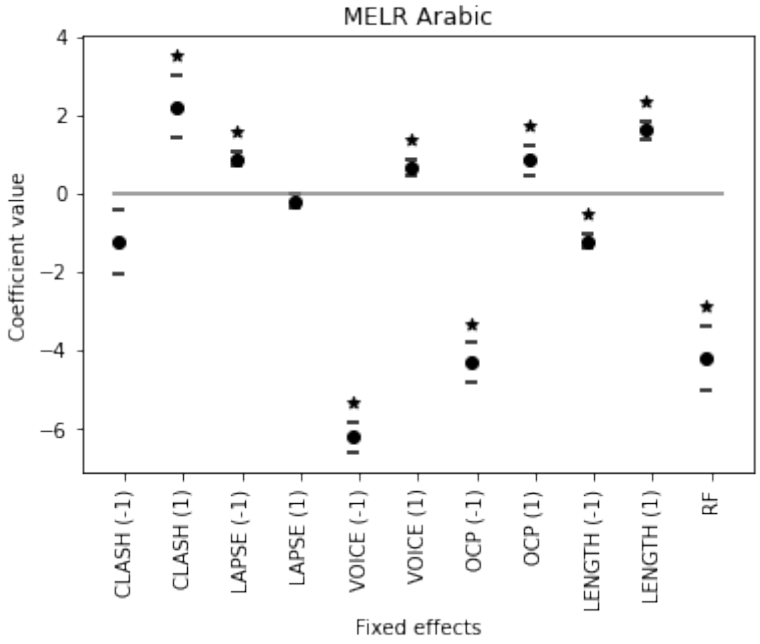


Figure 4.3: Visualization of the mixed-effects logistic regression results in Arabic. Coefficient values with standard error are shown for each fixed effect; significant effects are indicated by a star.

Discussion

No effect of Clash or Lapse, and significant effects of Voice, OCP, and Length on noun-adjective ordering in Arabic were expected, following the Phonologically-conditioned Hypothesis. Consistent effects of Voice, OCP, and

LENGTH were found that correspond to the avoidance of these marked sequences, as predicted.

A positive effect of CLASH (1) was found, indicating that postnominal clashes are avoided but not prenominal ones. Additionally, a positive effect of LAPSE (-1) on ordering was found, showing that it is significantly tolerated across the word boundary in postnominal order. The plurality of lapses (43%; 780/1805) are due to initial stress on word 1 with a length of three syllables or more, and non-initial stress on word 2, creating a sequence of 3 or more unstressed syllables across the boundary. Lapses are repaired in Palestinian Arabic (Houghton, 2008), but not necessarily Modern Standard Arabic, which is likely the Arabic comprising the majority of the Common Voice corpus. For this reason it was not expected to be avoided, and its tolerance in postnominal order may be due not only to the inactivity of LAPSE in the phonology of Arabic, but also to its stress system.

CHAPTER 5

DISCUSSION AND CONCLUSIONS

This chapter offers a larger discussion, contextualizing the results found across all five languages analyzed in this dissertation. At least one phonological constraint was found to influence noun-adjective ordering in every language, contributing to growing evidence that word ordering is conditioned by phonology cross-linguistically. A summary of the results is provided in section 5.1, and their implications for our understanding of the syntax-phonology interface are explored in section 5.2. Some limitations and future directions of this dissertation are discussed in 5.3. Section 5.4 concludes.

5.1 Results summary

A summary of the predictions for phonological constraints based on the Phonologically-conditioned Hypothesis and the results of the regression models across all languages is provided in Table 5.1. Constraints where the predictions of this hypothesis held true are: Length and OCP-Place in most languages; Hiatus in French and Italian; and Voice in French, Polish, and Arabic. Unexpected behavior was found for Length in French, Clash and Lapse in Italian, and Hiatus in Polish. A further discussion of these results and their implications are presented in the next section.

Constraint	Italian		French		Polish		Arabic	
Clash	✓	✗	–	–	–	–	✗	(✗)
Lapse	✓	✗	–	–	–	–	✗	(✗)
Hiatus	✗	✗	✓	✓	✓	✗	–	–
Voice	–	–	✓	(✓)	✓	(✓)	✓	✓
OCP-Place	–	–	✗	✗	✗	(✗)	✓	✓
Length	✓	✓	✓	✗	✓	(✓)	✓	✓

Table 5.1: Summary table of predictions (left) and results (right) in each language for the phonological constraints that are actively avoided (✓) via word-ordering or ignored/tolerated (✗). Parentheses in the table indicate constraints where only one of the simple effects supported the hypothesis.

5.2 Implications for phonology at its interfaces

5.2.1 Phonological nature of the effects

There has long been a discussion around what aspects of phonology are at the interface with syntax (Chomsky, 1965; Selkirk, 1978; Zwicky and Pullum, 1986; Zec and Inkelas, 1990, among others). Some have argued that only the prosodic hierarchy can interact with syntactic structure, under the prosodic view (Selkirk, 1978; Zec and Inkelas, 1990). Hayes (1990), however, points to three cases that show effects at the interface that are not strictly prosodic: tone in Ewe, liaison in French (discussed in this dissertation as well), and vowel length in Hausa (p.87).

In this dissertation, I found evidence for prosodic, syllabic, and segmental effects. The most consistent effect across languages is Length, which is the preference for short-before-long. Example (16) from French is repeated in (27).

(27) **Short before long preference**

a. un air avide
a air greedy
'a greedy air'

b. un avide hippopotame
a greedy hippopotamus
'a greedy hippopotamus'

LENGTH is a prosodic effect, as it deals with word length, here measured by syllable count. In many cases, there is no repair available for long-before-short sequences other than movement. Syllables cannot always be deleted from word 1 or added to word 2 as a repair in these languages. There are a few examples of word-length manipulation in order to satisfy prosodic well-formedness constraints. Anttila (2016) describes words in Mandarin that have long and short forms (two or one syllable) that can be manipulated to satisfy stress requirements in compounds, "with little syntactic or semantic difference" (p. 126). Similarly, verbs in English can be contracted only if the noun and the verb are both unstressed (Anttila, 2016). In any case, nouns and adjectives in these languages cannot be arbitrarily lengthened or shortened to satisfy heavy-final pressures.

Additional prosodic effects analyzed in this dissertation are CLASH and LAPSE, for which there is very little supporting evidence of their avoidance via word ordering. Recall that both of these constraints are argued to be active in Italian (Nespor and Vogel, 1979). CLASH is potentially phonologically repaired in Hindi, but there is not strong evidence (Pandey, 2021). Hindi results are inconclusive; however, models of the Italian data find no effects of either stress constraint. I argue that because clashes and lapses can be repaired in Italian at the word boundary with stress shift or RS, or beat addition (Nespor and Vogel, 1979, 1989), their syntactic repair is optional. Findings from the acoustic sample

of clash tolerances in Italian from the corpus suggest that phonological repairs are present only some of the time (38%), but may be found more consistently under experimental conditions (Burroni and Tilsen, 2022; Cohn and Renwick, 2021).

The HIATUS constraint targets syllable wellformedness, and was found to condition word ordering in French. Vowel hiatus is the occurrence of two vowels belonging to separate syllables without an intervening consonant. The cross-linguistic preference for CV syllables has been widely documented (Jakobson, 1962; Maddieson, 2013; Gordon, 2016), and this bias has been argued to exist for reasons of perception and production. Without an intervening consonant, syllable 2 in a V.V sequence has no onset and is therefore ill-formed.

The remaining two constraints: VOICE and OCP are segmental effects. VOICE was found to condition word ordering in French, Polish, and Arabic; and OCP in Arabic. Both of these constraints target aspects of segments, voicing and place of articulation specification. In the case of VOICE, the preference is for segments to agree in voicing: voiced-voiced or voiceless-voiceless. Voicing assimilation is almost always regressive (Lombardi, 1999), likely due to anticipatory articulation effects. In a typological study of onset clusters, Kreitman (2008) found that voiceless-voiceless clusters are the least marked cross-linguistically, and the presence of voiced-voiceless clusters in a language generally implies voiceless-voiced (implying then, of course, voiceless-voiceless). Voiced-voiced clusters also imply voiceless-voiceless clusters. The language-specific distinction between voiced-voiceless and voiceless-voiced consonant sequences across the word boundary is left to future work, as both were treated equally in the analysis presented in this dissertation. In the case of OCP, the preference is for seg-

ments to disagree in place of articulation Goldsmith (1976); McCarthy (1986). Converse to assimilatory effects, maintenance of place of articulation contrast is likely due to perceptual motivations, which is beyond the scope of this dissertation.

5.2.2 Impact of additional acoustic and semantic analyses

The main research question of this dissertation is whether word ordering is used to avoid markedness at the phonemic level, but answering this question would not be complete without a look at what surface forms are being produced and how the differences in meaning between orders may interact with phonology to affect order variation. Additional acoustic and semantic analyses for French and Italian were conducted to begin answering these questions.

Acoustic analyses of 50 tokens each in French and Italian offer further explanation for the phonological findings. In French, cases of noun-adjective pairs that did not avoid a violation of Voice where they could have showed that most of these were phonologically repaired by regressive voice assimilation or an *e muet* (at a combined rate of 70%, see section 2.4.2). Contrary to the hypotheses, Voice was not a significant predictor of ordering in the *similar* regression model in French, potentially due to the propensity of such repairs. In the *dissimilar* model, however, Voice (-1) was a significant predictor; rates of phonological repairs in this dataset and a further investigation into reasons for this discrepancy between models in French are left to future work. In Italian, cases of pairs that did not avoid Clash where they could have showed that most of these had no phonological repair (only 38%, see section 3.4.2). Clash was not a signif-

icant predictor in both models of Italian which, like VOICE in French, was not predicted. These regression results taken in combination with the results from the acoustic analysis contribute to a growing body of literature that reconsiders clash effects in Italian (e.g., Burroni and Tilsen, 2022; Burroni, 2022).

A larger, more comprehensive acoustic analysis of the realization of underlyingly marked phonological sequences would contribute to a greater understanding of the interaction between word ordering and phonology. I expect that such an analysis would reveal more variation, such as effects on additional acoustic variables like segment duration in French, or syllable duration or vowel formants in Italian. Including additional languages and phonological effects may reveal cross-linguistic differences and further test hypotheses about language-specific markedness. An examination of individual speaker behavior would be much better executed under a controlled experiment than with the corpora used in this dissertation.

In order to separate out some of the semantic effects on word ordering, the French and Italian datasets were each split into two based on the semantics of the adjective in the noun-adjective pair. Adjectives that had a larger cosine similarity value between the prenominal and postnominal embeddings were grouped into the *similar* dataset; those with a smaller value into the *dissimilar* dataset (see 1.4.3 for more details). Separate regression models were fit to noun-adjective pairs with an adjective that was *similar* between its positions, and to those with an adjective that was *dissimilar* between positions. It was predicted that phonological effects would be stronger or present only in the *similar* dataset, compared to the *dissimilar* one. There were no effects that were significantly greater in the similar models compared to the dissimilar. The SEMANTICALLY-CONDITIONED HY-

pothesis, which follows work by (Shih, 2014), is not supported by these results.

A more fine-grained or controlled semantic analysis may still reveal an interaction between phonological and semantic effects, predicted but not supported by this dissertation. Instead of two bins, *similar* and *dissimilar*, the dataset could be further subdivided, or a model with similarity as a continuous variable could be fit to a single dataset. Using controlled sets of adjectives known to be semantically neutral between positions and those known to be semantically different, a clearer comparison could be made of the presence or strength of phonological effects on these types of data.

Taken together, these two additional analyses provide a greater understanding of how phonology affects word ordering by looking at the acoustic realization of underlying markedness and the interaction of phonological effects with semantic ones.

5.2.3 Amendments to the original hypotheses

Given the findings presented in chapters 2 through 4, an amendment must be made to the Phonologically-conditioned Hypothesis, repeated below, from section 1.4.5.

Phonologically-conditioned Hypothesis: Only those phonologically-marked phenomena that are avoided with phonological repairs may also be avoided with syntactic repairs.

First, I want to highlight that the following specification in the original hypothesis holds true for the data presented in this dissertation: Phonological se-

quences avoided via word ordering must also be active in the language's phonology. There were no consistent cases of inactive phonological sequences having an effect on noun-adjective ordering: HIATUS in Italian, OCP in French and Polish, and the stress constraints in Arabic had at least one simple effect that was not significant or tolerated in the models where they are not avoided by the language-specific phonology.

Next, I want to propose, based on the results, that it is not true that *all* phonologically-active sequences are also avoided syntactically. Markedness in the (language-specific) phonology allows for the *option* of a syntactic repair. This is relevant for the VOICE, HIATUS (Polish), and the stress constraints. Polish and Arabic noun-adjective pairs avoided voice-disagreeing clusters; however, VOICE was a significant predictor in only the *dissimilar* model of French. Similarly, HIATUS was predicted to have an effect in Polish where it did not. In previous chapters, I have suggested that the lack of avoidance of these sequences via word ordering is due to the fact that they are phonologically reparable in either order; thus, it is possible to not have to resort to movement, which may be syntactically dispreferred or semantically costly. Results for the acoustic analysis of French suggest this may be the case. In the analysis of the sample of flexible noun-adjective pairs, 70% of tokens did not have a surface violation of VOICE: the cluster was repaired with assimilation or vowel epenthesis. An analysis of glottalization in Polish may reveal similar results. Lexical selection of a synonym as an avoidance strategy could be at play as well, in the case of VOICE in the *similar* French model or HIATUS in Polish (such a result was found in English by Breiss and Hayes (2020)); this effect is also left to future research.

Finally, I want to point out the distinction between phonologically-marked

sequences that are reparable at the word boundary, like voice assimilation, and those that are not. This notion, in tandem with the assertion that phonological effects higher in the prosodic hierarchy are relatively stronger, aligns with the consistency of the LENGTH effect across languages. It also may be the driving force behind the avoidance of HIATUS in French and OCP in Arabic. As detailed in section 2.2.1, hiatus is repaired in French by a process called *liaison*; however, this process is restricted to certain lexical items and environments (Tranel, 1995). Hiatus cannot always be repaired in a given noun-adjective ordering. Similarly, dissimilation processes that repair identical consonant sequences have been described to apply word-internally, and may not be available across the word boundary (Faiq and Burhanuddin, 2019). Such marked sequences that cannot be repaired phonologically may require avoidance via word ordering. A larger sample targeting the difference between constraints like HIATUS in French and OCP in Arabic and those like VOICE is an avenue for future research.

Given these facts, I propose a revised version of the hypothesis, below.

PHONOLOGICALLY-CONDITIONED HYPOTHESIS (REVISED): Only those phonologically-marked phenomena that are avoided with phonological repairs may *optionally* be avoided with syntactic repairs, if an alternative phonological repair is available.

Predictions of the SEMANTICALLY-CONDITIONED HYPOTHESIS were not borne out in the data. There were no effects that were present only in the similar models; nor were the effects present in both models statistically greater in the similar models. As previously noted, it may be the case that grouping the data into only two bins was not enough to minimize or target semantic differences. A model that further subdivides pairs based on semantics or one that uses only those adjectives that are known or confirmed by native speakers to have little

to no difference in meaning between orderings may be a better way to test this hypothesis. This more detailed examination of the interaction between semantic and phonological effects is a rich area of future research.

Semantically-conditioned Hypothesis: Phonological effects on ordering are stronger if semantic differences between orders are minimal.

5.2.4 The syntax-phonology interface

The principal contributions of this dissertation are methodological and empirical: I provide evidence that the surface ordering of noun-adjective pairs in various languages can, in part, be accounted for by the avoidance of phonological markedness such as vowel hiatus, clusters that disagree in voicing, violations of the OCP, and prosodic end weight. In this section, I discuss the implications of these findings on various theories about the nature of the syntax-phonology interface. Theories that consider linearization to be separate from syntax (Distributed Morphology, Halle et al., 1993; Holmberg, 1999) are not discussed as there is no possibility in these frameworks for phonology to condition syntax (Anttila, 2016).

Beginning most notably with Zwicky and Pullum (1986), the Principle of Phonology-Free Syntax (PPFS) dominated theories about the syntax-phonology interface for many years (Vogel and Kenesei, 1990; Miller et al., 1997, among others). In its original conception, the PPFS states that "no syntactic rule can be subject to language-particular phonological conditions or constraints" (Zwicky and Pullum, 1986, p.71). This theory emphasizes the distinction between the algebraic nature of the grammar, comprised of rules, and the statistical nature

of speaker behavior, referred to as tendencies. The grammar is composed of independent, distinct components which have their own rules[1]. Tendencies are considered to be outside of the grammar: they may be the result of diachrony, making them "accidents" from a synchronic perspective; of speaker preferences; or, of sociolinguistic variables. Under this view, the results presented in this dissertation are not an issue for PPFS because none of the phonological constraints on word ordering are absolute. For instance, producing the noun-adjective pair *maison magnifique* 'beautiful house' in the order in which the longer word is first, *magnifique maison*, does not render the phrase ungrammatical or unacceptable to a French speaker[2]. My results indicate that avoiding phonological markedness in sentence structure is preferred, but not categorical; as such, they constitute tendencies rather than grammatical rules and are not instances of phonologically-conditioned syntax under this definition.

Even taking non-categorical phonological effects on word ordering into account, some maintain that a theory of grammar where syntax is unaffected by phonology is still possible. In particular, Anttila (2016) argues for variation + filtering theory, in which variation generated by syntax is filtered by phonology. In this theory, the syntactic component of the grammar is responsible for the ordering of elements (hierarchical organization and linearization), and the choice of elements (lexical selection and morphology). Various linearizations are generated by the syntax; those that are phonologically ill-formed are ruled out, and those that are well-formed or not fatally ill-formed are allowed. See Figure 5.1.

This is formulated in an OT-style grammar (Prince and Smolensky, 2004),

[1]The issue of phonological conditioning on morphology is proposed to be part of a separate component, occurring after syntax and before phonology (Zwicky and Pullum, 1986, p.72).

[2]Such an effect, however, has been found elsewhere. Recall the Serbo-Croatian topicalization data from section 1.2 that shows that topicalized NPs containing a single phonological phrase are ungrammatical (Zec and Inkelas, 1990).

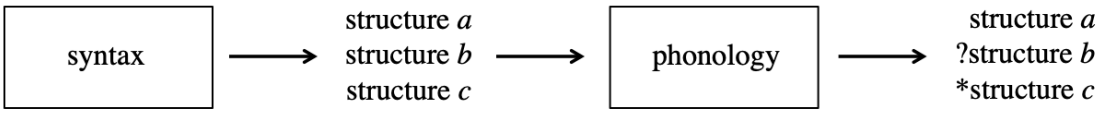


Figure 5.1: Conceptualization of variation + filtering theory, adapted from Anttila (2016).

in which the inputs are the linearization variants; syntactic constraints outrank alignment constraints that map syntactic structure to prosodic structure, which then outrank phonological constraints. In this organization, surface order is phonologically-conditioned, but a feed forward, syntax → phonology model of grammar, which interfaces at the level of prosody, is maintained. If the types of phonological constraints are expanded beyond prosody, my results fit well with this theory of the syntax-phonology interface. See an example of a variation + filtering grammar of French in Table 5.2, which selects the phonologically-optimal candidate *un important oiseau* /œ̃.nɛ̃.pɔʀ.tɑ̃.twa.zo/ 'an important bird' from multiple possible syntactic candidates. This candidate crucially does not violate HIATUS (compared to candidate *b* /œ̃.wa.zo.ɛ̃.pɔʀ.tɑ̃/).

		SYNTAX	ALIGNMENT	PHONOLOGY
a	☞ ([un important oiseau])			
b	([un oiseau important])			1
c	([un important)(oiseau])		1	
d	([un oiseau)(important])		1	1
e	([important un oiseau])	1		

Table 5.2: Example of variation + filtering theoretic grammar (Anttila, 2016). Syntactic constituents are indicated in square brackets and phonological constituents by parentheses.

In particular, some effects were found to be relevant to noun-adjective ordering in some languages but not others (e.g., HIATUS in French but not Polish),

which corresponds to the relative ranking of constraints in OT grammars that are language-specific (Prince and Smolensky, 2004). HIATUS is likely ranked higher in the grammar of French relative to that of Polish, given independent evidence of phonological processes (i.e., liaison); therefore, it would have a stronger conditioning effect on various linearizations generated by the syntax.

Towards the opposite end of the continuum from Zwicky and Pullum (1986) are those that argue for a bidirectional influence at the interface between syntax and phonology. Zec and Inkelas (1990) propose a theory of grammar wherein phonology can condition syntax with the restriction that it can only do so through prosody (Selkirk, 1978). They provide evidence from clitic ordering, word ordering, and topicalization which shows that sentences that do not obey prosodic rules are rendered ungrammatical. While prosodic constraints on cliticization, for example, may be achieved with a theory similar to variation + filtering, in the case of topicalization, Zec and Inkelas argue that both the syntactic component and the prosodic component must be simultaneously available. See example (7), reproduced here, wherein the topicalization in (28b) is ungrammatical due to prosody. A NP must be evaluated for topicalization at the same time that it is evaluated in terms of prosodic structure since only branching prosodic constituents of a certain syntactic kind can be topicalized.

(28) **Serbo-Croatian topicalization**

a. $[[\text{Petar}]_{\omega}\ [\text{Petrović}]_{\omega}]_{\text{NP}}$voleo-je Mariju
Peter Petrovic loved-AUX Mary
'Peter Petrovic loved Mary'

b. * $[[\text{Petar}]_{\omega}]_{\text{NP}}$voleo-je Mariju
Peter loved-AUX Mary
'Peter loved Mary'

Under the theory proposed by Zec and Inkelas (1990), syntax and phonology are mutually influential through the prosodic hierarchy; however, the results in this dissertation are not strictly limited to prosody, though effects of LENGTH were found. If the syntax-phonology interface is in fact bidirectional, it cannot be constrained solely to the prosodic component of phonology. Syllabic and segmental effects were found in this dissertation and other recent work (Breiss and Hayes, 2020; Shih and Zuraw, 2017).

Perhaps a better approach to characterizing the restrictions of the phonological conditions on syntax is that argued by Shih (2014), along with ideas put forth by Martin (2011). Constraints on phonological material have the potential to act on morphosyntactic material as well, but are generally weaker across greater constituent boundaries. For example, a phonological constraint against hiatus will be strongest within words, weaker across morpheme boundaries, and weaker still across word or phrase boundaries. This is argued to be a consequence of learning: speakers learn phonotactic or phonological constraints and overgeneralize them to greater boundaries (Martin, 2011). Phonological constraints on morphosyntax must also compete with semantic, syntactic, and usage-based factors under this view, which additionally contributes to their weakening (Shih, 2014). See Figure 5.2 for a visualization of the relative strength of phonological constraints, which are strongest within words at the left edge of the figure and weakest across phrases at the right edge, where they compete with other factors[3].

These two characteristics of phonological constraints on syntax, that they weaken across boundaries and with the presence of non-phonological effects,

[3]Prosodic-edge effects, like domain-initial strengthening (e.g., Keating et al., 2004) and domain-final devoicing (e.g., Iverson and Salmons, 2007), however, have been argued in favor of phonological *strengthening* at greater boundaries.

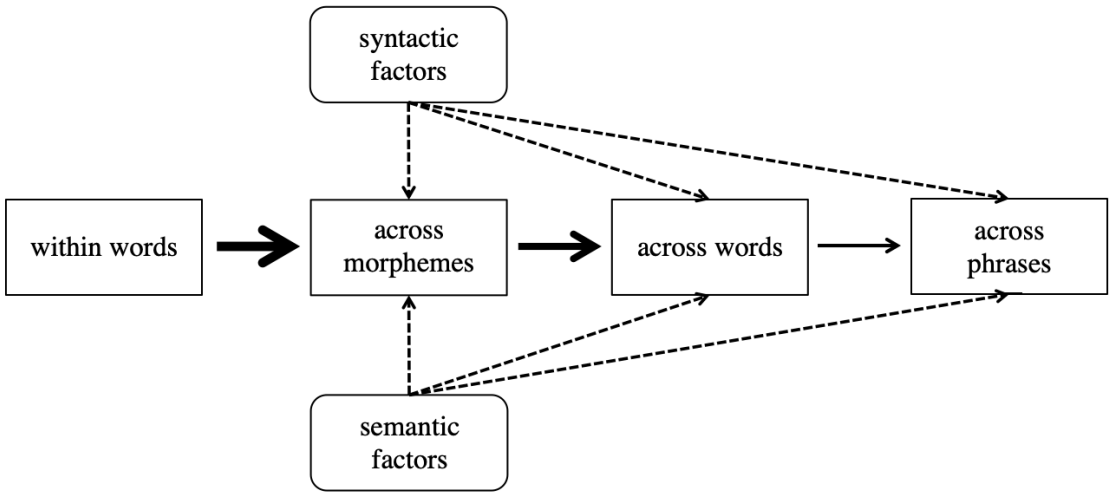


Figure 5.2: Conceptualization of the relative strength of phonological effects, which have been theorized to weaken across boundaries and with the presence of non-phonological constraints (Shih, 2014; Martin, 2011).

align well with the results presented in this dissertation. Phonological effects not active at the level of the word were also not found to be significant predictors of word ordering, such as OCP in French; and those that are categorical at word level are gradient at the level of the word boundary, such as VOICE in Polish. It was not confirmed by this work, however, that when semantic effects have a strong co-presence, phonological effects are weaker or not present. This hypothesis may be supported by more nuanced work in the future.

5.3 Limitations and future directions

This dissertation is not without its limitations, which are noted here as important caveats, but also in the hopes that they will help inform future work. A richer corpus, in a few respects, would contribute to a deeper understanding of

phonological effects on word ordering. First, these corpora come from Common Voice and lack sufficient dialect information. Degree of flexibility and strength or presence of phonological effects on noun-adjective pairs could vary greatly between dialects. A by-dialect analysis may reveal stronger or different phonological effects in some dialects, or weaker or no effects in others; the dialectal variation present in the dataset may be obscuring results (Cohn and Renwick, 2021). These corpora also are built by the small contributions of many, rather than the large contributions of few, which makes the role of inter- versus intra-speaker variation difficult to ascertain, and again may obscure effects. Some speakers may have greater noun-adjective flexibility or a stronger preference for avoiding phonological markedness. Finally, knowing or controlling for the degree of formality in speech used is also likely to contribute to a clearer understanding of these effects. In addition to the corpora used, the method of using phonemic forms from lexical databases (French, Italian) or grapheme to phoneme methods (Polish, Hindi, Arabic) is less than ideal. From a theoretical perspective, it is not certain whether these forms are actually the psychological reality for all speakers included in the corpus. Computationally, a lot of noun-adjective pair data had to be excluded due to missing phonological forms using these methods, especially in Polish and Hindi. Again, results may be obscured due to missing data. Finally, relative frequency was calculated over the frequencies of pairs found in the same corpus used for analysis. Given the limitations of these corpora, including dialect and register, there may be flexible noun-adjective pairs that are not present in the data or have a different frequency distribution.

There are several future areas of research, in addition to the methodological improvements noted above and other suggestions throughout this dissertation. In this work, I began to build a typology of the phonological markedness effects

on noun-adjective ordering. Three of the languages analyzed have postnominal NOUN ADJECTIVE dominant order (French, Italian, and Arabic), and two have prenominal ADJECTIVE NOUN dominant order (Polish and Hindi). The typological differences between these two types of languages may be further explored in future work, especially as it bears on different syntactic and semantic theories of adjectives, described below.

Though this dissertation assumed the syntactic and semantic properties of adjectives as outlined by Cinque (2010), there are alternative approaches that could be explicitly tested in future work. Alexiadou (2014) outlines two traditional schools of thought about adjective placement: those ascribing to the 'separatist' approach, wherein the adjective is base-generated in a language-specific position (e.g., Cinque, 1993; Sproat and Shih, 1988) and those promoting the 'reductionism' approach, wherein the adjective is universally base-generated in a single position and moves according to language-specific rules resulting in different surface structures (e.g., Jacobs and Rosenbaum, 1968; Kayne, 1994; Cinque, 2010). Under a separatist approach, movement of an adjective to a non-dominant position to avoid phonological markedness would cost the same in prenominal-dominant and postnominal-dominant languages. Under the reductionist approach, where adjectives are base-generated in postnominal position regardless of the language-specific dominant surface order, movement to a non-dominant position may be more costly in a postnominal-dominant language compared to a prenominal-dominant one. This dissertation did not explicitly test these two syntactic different hypotheses, but the data from the three postnominal languages in comparison to the two prenominal languages could be reexamined in future work to look at this issue in greater depth.

The relative ordering of adjective and noun does not solely affect those two elements; there are often other words ("neighbors") in the utterance at the word edge of the adjective and noun that could be affected by their ordering. In previous work, I looked at clash between adjectives and nouns and their neighbors, comparing it to rates of clash at the noun-adjective word boundary, where the adjective was flexible. Clash was tolerated at a higher rate at the boundaries with a neighbor, than at the noun-adjective boundary. Future work could explore this distinction further, or extend the types of boundaries investigated to include subject, verb, and object, which would expand this work to phonological phrases.

Though it was not the focus of this dissertation, a reexamination of these data could test hypotheses related to the emergence of the unmarked (TETU), an important tenet of OT which states that though constraints may be crucially dominated in a language, they are never "turned off" (McCarthy and Prince, 1994). Theoretically, this predicts that there exist places in the grammar where evidence of these constraints can be observed, where the dominating constraints are not relevant. From this, one would expect that a constraint that is not active in a language's grammar, such as OCP in French, would still be avoided in non-default word order because switching to default word order would only increase the phonological and syntactic wellformedness of the phrase. This was not observed in the results of this dissertation. Each constraint is comprised of two simple effects: Constraint (-1), which shows the correlation of wellformedness in postnominal order with prenominal or postnominal order; and Constraint (1), which shows the correlation of wellformedness in prenominal order with prenominal or postnominal order. Support for TETU may hypothetically be found in inactive constraints that are well-formed only in a language's default

order (thus there may be less of a syntactic cost to surfacing in the default order, which also avoids phonological markedness, though there may still be a semantic cost). There were only two simple effects found to be significant in languages where they are hypothesized to be inactive: OCP (-1) in Polish and CLASH (1) in Arabic; however, both of these were cases of using *non*-default word ordering to avoid a violation of the constraint in default order. TETU should theoretically manifest in utterances in which the phonologically-inactive constraint is the only constraint that could be violated in non-default order; otherwise, it may be the case that the pressure to avoid the violation of a phonologically-*active* constraint is the driving force behind the word ordering. Future work could isolate these cases and see if there is indeed evidence for the emergence of the unmarked in noun-adjective ordering.

The processing and encoding of phonological effects on word ordering also poses many compelling questions. An examination of the role of diachrony may reveal additional phonological effects not currently active in the synchronic grammar. Noun-adjective pair construction with inflexible adjectives may also tend to avoid phonological markedness, possibly to a different degree than what is observed here. The degree to which the effects found in this dissertation are produced online or are encoded in the grammar of the speaker is a ripe avenue of future research.

5.4 Conclusions

This dissertation provided evidence from five languages that word ordering is phonologically conditioned. This effect is shown via the investigation of various

phonological effects on the ordering of nouns and adjectives, where the adjective has flexible placement. The phonological constraints examined in this work dealt with: prosody, the length of words and word-level stress; syllable structure, the preference for onsets; and segmental effects of voicing and place of articulation. Results show that all three types of phonological markedness may be avoided at the word boundary by a preference for an order where a violation does not surface. The most consistent effect across languages was that of LENGTH, a preference for shorter before longer words.

The principal hypotheses of this work more or less held true. The constraints avoided in word ordering were all a subset of the constraints that are active in the language-specific phonology. Constraints that are phonologically active may be optionally avoided via word ordering, as seen in the case of stress constraints in Italian and the VOICE constraint in French. The supporting evidence for both of these hypotheses confirms previous work on the nature of the syntax-phonology interface, in particular that of Shih (2014) and Martin (2011).

This dissertation also presented a preliminary investigation of the acoustic reality of cases where violations of phonological constraints were tolerated where they could have been avoided via word ordering. The analysis of these samples of data in French and Italian revealed that phonological repairs for the violations were not always present. These results highlight differences between corpus and experimental work, and the in-depth investigation of speaker production and perception of these phenomena are an avenue of future research.

This work also has important methodological contributions. I developed an analysis pipeline that is generalized across languages, phonological constraints, and part-of-speech sequences. My publicly available scripts extract sentences

from Common Voice corpus files, tag them for parts of speech using an open-source tagger, subset the data for target sequences (such as noun-adjective pairs in this work), add phonological information from a lexicon, and code target sequences for order preferences based on phonological constraints. The semantic clustering script is also available, in addition to other analysis and supporting scripts.

In sum, this dissertation provided empirical evidence for phonological effects on surface word order, showing that language-specific phonological markedness helps determine the presence of these effects.

REFERENCES

Abeillé, A. and Godard, D. (1999). La position de l'adjectif épithète en français: le poids des mots. *Recherches linguistiques de Vincennes*, (28):9–32.

Abraham, R. D. (1950). Fixed order of coordinates: A study in comparative lexicography. *The Modern Language Journal*, 34(4):276–287.

Abu-Mansour, M. H. (1996). Voice as a privative feature: Assimilation in Arabic. *Amsterdam Studies in the Theory and History of Linguistic Science Series 4*, pages 201–232.

Adjekum, G., Holman, M. E., and Holman, T. W. (1993). *Phonological processes in Anufo*. Institute of African Studies, University of Ghana.

Alexiadou, A. (2014). The syntax of adjectives. In *The Routledge handbook of syntax*, pages 107–125. Routledge.

Allan, K. (1987). Hierarchies and the choice of left conjuncts (with particular attention to English). *Journal of Linguistics*, 23(1):51–77.

Altakhaineh, A. R. M. and Zibin, A. (2014). Phonologically conditioned morphological process in Modern Standard Arabic: An analysis of Al-ibdal 'substitution' in ftaʕal pattern using prosodic morphology. *International Journal of English Language and Linguistics Research*, 2(1):1–16.

Anttila, A. (2016). Phonological effects on syntactic variation. *Annual Review of Linguistics*, 2:115–137.

Assaneo, M. F. and Poeppel, D. (2018). The coupling between auditory and motor cortices is rate-restricted: Evidence for an intrinsic speech-motor rhythm. *Science advances*, 4(2):eaao3842.

Baković, E. (2005). Antigemination, assimilation and the determination of identity. *Phonology*, 22(3):279–315.

Benor, S. B. and Levy, R. (2006). The chicken or the egg? A probabilistic analysis of English binomials. *Language*, 82(2):233–278.

Berri, A. (2006). Aspects phonétiques et phonologiques du e-muet du français. *Fragmentos: Revista de Língua e Literatura Estrangeiras*, 30.

Berruto, G. (1987). *Sociolinguistica dell'italiano contemporaneo*, volume 33. Carocci.

Bhat, R. A., Bhatt, R., Farudi, A., Klassen, P., Narasimhan, B., Palmer, M., Rambow, O., Sharma, D. M., Vaidya, A., Vishnu, S. R., et al. (2017). The Hindi/Urdu treebank project. In *Handbook of Linguistic Annotation*. Springer Press.

Bing, J. M. (1980). Linguistic rhythm and grammatical structure in Afghan Persian. *Linguistic Inquiry*, pages 437–463.

Bloomfield, L. (1933). Language. *New York, Holt, Rinehart et Winston*.

Boersma, P. and Weenik, D. (2022). Praat: doing phonetics by computer. *http://www. praat. org/*.

Bolinger, D. L. (1962). Binomials and pitch accent. *Lingua*, 11:34–44.

Bosco, C., Simonetta, M., and Maria, S. (2013). Converting Italian treebanks: Towards an Italian Stanford Dependency Treebank. In *7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 61–69. The Association for Computational Linguistics.

Breiss, C. and Hayes, B. (2020). Phonological markedness effects in sentence formation. *Language*, 96(2):338–370.

Bresnan, J., Cueni, A., Nikitina, T., and Baayen, R. H. (2007). Predicting the dative alternation. In *Cognitive foundations of interpretation*, pages 69–94. KNAW.

Buchanan, K. N. (2012). *Perspectives on quantity-sensitivity and decomposed scalar constraints: A view from Hindi stress*. PhD thesis, UC Santa Cruz.

Buckwalter, T. (2004). Issues in Arabic orthography and morphology analysis. In *Proceedings of the workshop on computational approaches to Arabic script-based languages*, pages 31–34.

Bullinaria, J. A. and Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior research methods*, 39(3):510–526.

Burroni, F. (2022). A split-gesture, competitive, coupled oscillator model of syllable structure predicts the emergence of edge gemination and degemination. *Proceedings of the Society for Computation in Linguistics*, 5(1):11–22.

Burroni, F. and Tilsen, S. (2022). The online effect of clash is durational lengthening, not prominence shift: Evidence from Italian. *Journal of Phonetics*, 91:101124.

Candito, M. and Seddah, D. (2012). Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. In *TALN 2012 - 19e conférence sur le Traitement Automatique des Langues Naturelles*, Grenoble, France.

Cardinaletti, A. (2010). On a (wh-) moved topic in Italian, compared to germanic. *Advances in comparative Germanic syntax*, pages 3–40.

Casali, R. F. (2021). *Resolving hiatus*. Routledge.

Chomsky, N. (1965). Aspects of the theory of syntax.

Cinque, G. (1993). A null theory of phrase and compound stress. *Linguistic inquiry*, 24(2):239–297.

Cinque, G. (2010). *The syntax of adjectives: A comparative study*, volume 57. MIT press.

Clogg, C. C., Petkova, E., and Haritou, A. (1995). Statistical methods for comparing regression coefficients between models. *American Journal of Sociology*, 100(5):1261–1293.

Cohn, A. C. and Renwick, M. E. (2021). Embracing multidimensionality in phonological analysis. *The Linguistic Review*, 38(1):101–139.

Cooper, W. E. and Ross, J. R. (1975). World order. *Papers from the parasession on functionalism*, pages 63–111.

Côté, M.-H. (1997). Phonetic salience and the OCP in coda cluster reduction. In *Meeting of the Chicago Linguistic Society*, volume 33, pages 57–71.

Coupé, C., Oh, Y. M., Dediu, D., and Pellegrino, F. (2019). Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche. *Science advances*, 5(9):eaaw2594.

Dell, F. (1984). L'accentuation dans les phrases en français. *Forme sonore du langage*, pages 65–122.

Dell, F. et al. (1980). *Generative phonology and French phonology*. CUP Archive.

Dryer, M. S. and Haspelmath, M., editors (2013). *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Durand, J. and Lyche, C. (2008). French liaison in the light of corpus data. *Journal of French Language Studies*, 18(1):33–66.

El-Ayoubi, H., Ayyūbī, H. I., Fischer, W., and Langer, M. (2001). *Syntax der arabischen Schriftsprache der Gegenwart*, volume 2. Reichert Verlag.

Faiq, A. M. and Burhanuddin, I. (2019). The process of dissimilation in English and Arabic: A comparative study. *Journal of Al-Frahedis Arts*|مجلة آداب الفراهيدي, (29):320–337.

Fenk-Oczlon, G. (1989). Word frequency and word order in freezes.

Forsgren, M. (1978). *La place de l'adjectif épithète en français contemporain: étude quantitative et sémantique*. PhD thesis, Acta Universitatis Upsaliensis.

Fougeron, C. and Smith, C. (1993). Illustrations of the ipa: French. *Journal of the International Phonetic Association*, 23(2):73–76.

Fournier, R. (1978). De quelques anomalies dans le traitement de l'article défini par H. Tinelli (1970): Generative phonology of Haitian Creole. *Amsterdam Creole Studies*, 2:101–115.

Frisch, S. A. and Zawaydeh, B. A. (2001). The psychological reality of OCP-Place in Arabic. *Language*, pages 91–106.

Goldsmith, J. A. (1976). *Autosegmental phonology*. PhD thesis, Massachusetts Institute of Technology.

Golenbock, J. (2000). Binomial expressions—does frequency matter. *Unpublished ms., Carnegie Mellon University*.

Gordon, M. (1999). *Syllable weight: phonetics, phonology, and typology*. PhD thesis, University of California, Los Angeles.

Gordon, M. K. (2016). *Phonological typology*, volume 1. Oxford University Press.

Goslin, J., Galluzzi, C., and Romani, C. (2014). PhonItalia: a phonological lexicon for Italian. *Behavior research methods*, 46(3):872–886.

Green, T. and Kenstowicz, M. (1995). The lapse constraint.

Greenberg, J. H. (1950). The patterning of root morphemes in Semitic. *Word*, 6(2):162–181.

Griffiths, J. M. et al. (2020). *A quantitative reanalysis of schwa realization in contemporary metropolitan French*. PhD thesis.

Gunkel, D. C. and Ryan, K. (2011). Hiatus avoidance and metrification in the Rigveda.

Gussmann, E. (1992). Resyllabification and delinking: The case of Polish voicing. *Linguistic Inquiry*, 23(1):29–56.

Gustafsson, M. (1974). The phonetic length of the members in present-day English binomials. *Neuphilologische mitteilungen*, pages 663–677.

Hahn, M., Degen, J., Goodman, N. D., Jurafsky, D., and Futrell, R. (2018). An information-theoretic explanation of adjective ordering preferences. In *CogSci*.

Hall, R. A. (1948). *Descriptive Italian Grammar*, volume 2. Cornell University Press and Linguistic Society of America.

Halle, M., Marantz, A., Hale, K., and Keyser, S. J. (1993). Distributed morphology and the pieces of inflection. *1993*, pages 111–176.

Hallé, P. and Adda-Decker, M. (2007). Voicing assimilation in journalistic speech. In *16th International congress of phonetic sciences*, volume 2007, pages 493–496.

Hayes, B. (1980). *A metrical theory of stress rules.* PhD thesis, Massachusetts Institute of Technology.

Hayes, B. (1984). The phonology of rhythm in English. *Linguistic Inquiry*, 15(1):33–74.

Hayes, B. (1989). Compensatory lengthening in moraic phonology. *Linguistic inquiry*, 20(2):253–306.

Hayes, B. (1990). Precompiled phrasal phonology. *The phonology-syntax connection*, 85:108.

Hayes, B. and Puppel, S. (2019). On the rhythm rule in Polish. In *Advances in nonlinear phonology*, pages 59–82. De Gruyter Mouton.

Holmberg, A. (1999). Remarks on Holmberg's generalization. *Studia linguistica*, 53(1):1–39.

Houghton, P. (2008). Positionally licensed extended lapses. *University of Pennsylvania Working Papers in Linguistics*, 14(1):16.

Hume, E. (2011). Markedness. *The Blackwell Companion to Phonology*, 1:79–106.

Huszthy, B. (2016). Italian as a voice language without voice assimilation. *Proceedings of ConSOLE XXIV*, 428:452.

Hyman, L. (1985). A theory of phonological weight.

Inkelas, S. and Zec, D. (1995). Syntax-phonology interface. The Handbook of Phonological Theory, edited by John Goldsmith, 535-549.

Ishihara, S. (2015). Syntax-phonology interface. In *Handbook of Japanese phonetics and phonology*, pages 569–618. De Gruyter Mouton.

Iverson, G. K. and Salmons, J. C. (2007). Domains and directionality in the evolution of German final fortition. *Phonology*, 24(1):121–145.

Jacobs, R. A. and Rosenbaum, P. S. (1968). English transformational grammar.

Jain, U. R. (1995). *Introduction to Hindi grammar*. Center for South & Southeast.

Jakobson, R. (1962). Selected writings. vol. 1, phonological studies.

Kabrah, R., Broselow, E., and Ouali, H. (2011). Regressive voicing assimilation in Cairene Arabic. *Perspectives on Arabic Linguistics XXII–XXIII*, pages 21–33.

Kachru, Y. (2006). *Hindi*, volume 12. John Benjamins Publishing.

Kadenge, M. (2013). Hiatus resolution in Nambya: an Optimality Theory analysis. *Language Matters*, 44(1):94–121.

Kamusella, T. (2017). The arabic language: A latin of modernity? *Journal of Nationalism, Memory & Language Politics*, 11(2):117–145.

Kang, H.-S. (1998). The deletion of the glide y in Seoul Korean: Toward its explanations. 어학연구.

Kayne, R. S. (1994). *The antisymmetry of syntax*, volume 25. MIT press.

Keating, P., Cho, T., Fougeron, C., and Hsu, C.-S. (2004). Domain-initial articulatory strengthening in four languages. *Phonetic interpretation: Papers in Laboratory Phonology*, VI:143–161.

Kelkar, A. R. (1968). *Studies in Hindi-Urdu, Introduction & Word Phonology*. Deccan College Postgraduate and Research Institute.

Kenstowicz, M. J. (1994). *Phonology in generative grammar*, volume 7. Blackwell Cambridge, MA.

Kimball, J. (1973). Seven principles of surface structure parsing in natural language. *Cognition*, 2(1):15–47.

Kiparsky, P. (1966). Über den deutschen Akzent. *Studia grammatica*, 7:69–98.

Knittel, M. L. (2005). Some remarks on adjective placement in the French NP. *Probus*, 17(2):185–226.

Kramer, M. (2009). *The phonology of Italian*. Oxford University Press on Demand.

Kreitman, R. (2008). The phonetics and phonology of onset clusters: The case of Modern Hebrew.

Kremers, J. (2003). *The Arabic noun phrase. Netherlands Graduate School of Linguistics*. PhD thesis, Dissertation.

Laenzlinger, C. (2005). French adjective ordering: Perspectives on DP-internal movement types. *Lingua*, 115(5):645–689.

Lahousse, K. and Lamiroy, B. (2012). Word order in French, Spanish and Italian: A grammaticalization account. *Folia linguistica*, 46(2):387–416.

Lance, D. (1968). *Sequential ordering in prenominal modifiers in English: a critical review.* PhD thesis, University of Texas, Austin.

Leben, W. R. (1973). *Suprasegmental phonology.* PhD thesis, Massachusetts Institute of Technology.

Lee, J. L., Ashby, L. F., Garza, M. E., Lee-Sikka, Y., Miller, S., Wong, A., McCarthy, A. D., and Gorman, K. (2020). Massively multilingual pronunciation modeling with WikiPron. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4223–4228, Marseille, France. European Language Resources Association.

Levy, O. and Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. *Advances in neural information processing systems*, 27.

Levy, O., Goldberg, Y., and Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the association for computational linguistics*, 3:211–225.

Liberman, M. and Prince, A. (1977). On stress and linguistic rhythm. *Linguistic inquiry*, 8(2):249–336.

Liberman, M. Y. (1975). *The intonational system of English.* PhD thesis, Massachusetts Institute of Technology.

Lin, H.-s. (2018). Vowel hiatus resolution in Kavalan. *Taiwan Journal of Linguistics*, 16(1):53–93.

Local, T. (2014). French language is on the up, report reveals.

Lombardi, L. (1999). Positional faithfulness and voicing assimilation in Optimality Theory. *Natural Language & Linguistic Theory*, 17(2):267–302.

Lucci, V. (1976). Le mécanisme du "E" muet dans différentes formes de français parlé. *La Linguistique*, 12(Fasc. 2):87–104.

Maddieson, I. (2013). Syllable structure. *The world atlas of language structures online*.

Maiden, M. and Robustelli, C. (2014). *A reference grammar of modern Italian*. Routledge.

Malisz, Z., Żygis, M., and Pompino-Marschall, B. (2013). Rhythmic structure effects on glottalisation: A study of different speech styles in Polish and German. *Laboratory Phonology*, 4(1):119–158.

Malkiel, Y. (1959). Studies in irreversible binomials. *Lingua*, 8:113–160.

Martin, A. (2011). Grammars leak: Modeling how phonotactic generalizations interact within the grammar. *Language*, 87(4):751–770.

McCarthy, J. J. (1986). OCP effects: Gemination and antigemination. *Linguistic inquiry*, 17(2):207–263.

McCarthy, J. J. (1994). The phonetics and phonology of Semitic pharyngeals. *Papers in laboratory phonology III: Phonological structure and phonetic form*, 86:191–233.

McCarthy, J. J. (2018). *Formal problems in Semitic phonology and morphology*, volume 17. Routledge.

McCarthy, J. J. and Prince, A. (1994). The Emergence of the Unmarked: Optimality in Prosodic Morphology.

McDonald, J. L., Bock, K., and Kelly, M. H. (1993). Word and world order: Semantic, phonological, and metrical determinants of serial position. *Cognitive psychology*, 25(2):188–230.

McGregor, R. S. (1977). *Outline of Hindi Grammar*. Oxford University Press, Delhi. 2nd edition.

Meinschaefer, J. (2005). The prosodic domain of Italian troncamento is not the clitic group.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Miller, P. H., Pullum, G. K., and Zwicky, A. M. (1997). The principle of phonology-free syntax: four apparent counterexamples in French. *Journal of Linguistics*, 33(1):67–90.

Mohamed, E. (2014). Object and subject Heavy-NP shift in Arabic.

Mollin, S. (2012). Revisiting binomial order in English: ordering constraints and reversibility. *English Language & Linguistics*, 16(1):81–103.

Mollin, S. (2013). Pathways of change in the diachronic development of binomial reversibility in Late Modern American English. *Journal of English Linguistics*, 41(2):168–203.

Morales, A. (1995). On deletion rules in Catalan. *AMSTERDAM STUDIES IN THE THEORY AND HISTORY OF LINGUISTIC SCIENCE SERIES 4*, pages 37–37.

Morgan, E. I. P. (2016). *Generative and item-specific knowledge of language*. University of California, San Diego.

Morin, Y. C. (2011). Remarks on prenominal liaison consonants in French. In *Living on the Edge*, pages 385–400. De Gruyter Mouton.

Nespor, M. (2019). The Phonological Word in Italian. *Advances in Nonlinear Phonology*, 7:193.

Nespor, M. and Vogel, I. (1979). Clash avoidance in Italian. *Linguistic Inquiry*, 10(3):467–482.

Nespor, M. and Vogel, I. (1986). *Prosodic Phonology*.

Nespor, M. and Vogel, I. (1989). On clashes and lapses. *Phonology*, 6(1):69–116.

New, B., Pallier, C., Brysbaert, M., and Ferrand, L. (2004). Lexique 2: A new French lexical database. *Behavior Research Methods, Instruments, & Computers*, 36(3):516–524.

Newlin-Łukowicz, L. (2012). Polish stress: looking for phonetic evidence of a bidirectional system. *Phonology*, 29(2):271–329.

Nølke, H. (1996). Où placer l'adjectif épithète? Focalisation et modularité. *Langue française*, pages 38–58.

Nothman, J., Ringland, N., Radford, W., Murphy, T., and Curran, J. R. (2017). Learning multilingual named entity recognition from wikipedia.

Ohala, M. (1983). *ASPECTS OF HINDI PHONOLOGY*, volume 2. Motilal Banarsidass Publisher.

Ohala, M. (2001). Some patterns of unscripted speech in Hindi. *Journal of the International Phonetic Association*, 31(1):115–126.

Pandey, P. (2021). An Optimality Theoretic account of word stress in Hindi. *Lingua*, 250:102994.

Pasha, A., Al-Badrashiny, M., Diab, M. T., El Kholy, A., Eskander, R., Habash, N., Pooleery, M., Rambow, O., and Roth, R. (2014). Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic. In *Lrec*, volume 14, pages 1094–1101. Citeseer.

Pierrehumbert, J. (1993). Dissimilarity in the Arabic verbal roots. In *Proceedings of NELS*, volume 23, pages 367–381. University of Massachusetts.

Pinker, S. and Birdsong, D. (1979). Speakers' sensitivity to rules of frozen word order. *Journal of Verbal Learning and Verbal Behavior*, 18(4):497–508.

Prince, A. (1990). Quantitative consequences of rhythmic organization. *Cls*, 26(2):355–398.

Prince, A. and Smolensky, P. (2004). *Optimality Theory: Constraint interaction in generative grammar*. John Wiley & Sons.

Prince, A. S. (1975). *The phonology and morphology of Tiberian Hebrew.* PhD thesis, Massachusetts Institute of Technology.

Prince, A. S. (1983). Relating to the grid. *Linguistic inquiry*, pages 19–100.

Przepiórkowski, A. (2012). *Narodowy korpus języka polskiego*. Naukowe PWN.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Quirk, R., Greenbaum, S., Leech, G. N., Svartvik, J., et al. (1972). A grammar of contemporary English.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rice, K. (2007). Markedness in phonology. *The Cambridge handbook of phonology*, pages 79–97.

Rischel, J. (1972). Compound stress in Danish without a cycle. *Annual Report of the Institute of Phonetics, University of Copenhagen No*, 6:211–218.

Rubach, J. and Booij, G. E. (1985). A grid theory of stress in Polish. *Lingua*, 66(4):281–320.

Ryan, K. M. (2019). Prosodic end-weight reflects phrasal stress. *Natural Language & Linguistic Theory*, 37(1):315–356.

Ryding, K. C. (2005). *A reference grammar of modern standard Arabic*. Cambridge University Press.

Sadowska, I. (2012). *Polish: A comprehensive grammar*. Routledge.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523.

Sapir, E. (1921). An introduction to the study of speech. *Language*, 1.

Schlüter, J. (2005). *Rhythmic Grammar: The Influence of Rhythm on Grammatical Variation and Change in English*. Walter de Gruyter.

Schlüter, J. and Knappe, G. (2018). Synonym selection as a strategy of stress clash avoidance. In *Corpora and lexis*, pages 69–105. Brill Rodopi.

Schütze, H. (1993). Word space. *Advances in neural information processing systems*, 5.

Schwartz, G. (2013). Vowel hiatus at Polish word boundaries–phonetic realization and phonological implications. *Poznań Studies in Contemporary Linguistics*, 49(4):557–585.

Scontras, G., Degen, J., and Goodman, N. D. (2019). On the grammatical source of adjective ordering preferences. *Semantics and Pragmatics*, 12:7.

Selkirk, E. (1978). On prosodic structure and its relation to syntactic structure. *Indiana University Linguistics Club*, 11:563–605.

Selkirk, E. (1984). Phonology and syntax: the relation between sound and structure.

Selkirk, E. et al. (2011). The syntax-phonology interface. *The handbook of phonological theory*, 2:435–483.

Selkirk, E. O. (1980). *On prosodic structure and its relation to syntactic structure*, volume 194. Indiana University Linguistics Club.

Shih, S. S. (2017). Phonological influences in syntactic alternations. *The morphosyntax-phonology connection: Locality and directionality at the interface*, pages 223–252.

Shih, S. S. and Zuraw, K. (2017). Phonological conditions on variable adjective and noun word order in Tagalog. *Language*, 93(4):e317–e352.

Shih, S. S.-y. (2014). *Towards optimal rhythm*. PhD thesis, Stanford University.

Siewierska, A. (1993). Syntactic weight vs information structure and word order variation in Polish. *Journal of Linguistics*, 29(2):233–265.

Singh, S. and Sarma, V. M. (2011). Verbal inflection in Hindi: A distributed morphology approach. In *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation*, pages 283–292.

Siyanova-Chanturia, A., Conklin, K., Caffarra, S., Kaan, E., and van Heuven, W. J. (2017). Representation and processing of multi-word expressions in the brain. *Brain and language*, 175:111–122.

Snoeren, N. D. and Segui, J. (2003). A voice for the voiceless: Voice assimilation in French. In *Proceedings of the 15th International Congress of Phonetic Sciences*, pages 2325–2328. Citeseer.

Speyer, A. (2008). *Topicalization and clash avoidance. On the interaction of prosody and syntax in the history of English with a few spotlights on German*. University of Pennsylvania.

Spitzer, L. (1941). French pataquès; cuir. *Language*, pages 253–255.

Sproat, R. and Shih, C. (1988). Prenominal adjectival ordering in English and Chinese. In *Proceedings of NELS*, volume 18, pages 456–489.

Sproat, R. and Shih, C. (1991). The cross-linguistic distribution of adjective ordering restrictions. In *Interdisciplinary approaches to language*, pages 565–593. Springer.

Stowell, T. (1979). Stress systems of the world, unite! *MIT Working Papers in Linguistics*, 1:51–76.

Svantesson, J.-O., Tsendina, A., Karlsson, A., and Franzén, V. (2005). *The phonology of Mongolian*. Oxford University Press.

Swan, O. E. (2002). *A grammar of contemporary Polish*. Slavica, Bloomington, Indiana.

Tamim, N. (2017). *Voicing contrast of stops in the Palestinian Arabic dialect*. PhD thesis, Universiteit Van Amsterdam.

Thuilier, J. (2012). *Contraintes préférentielles et ordre des mots en français*. PhD thesis, Université Paris-Diderot-Paris VII.

Tranel, B. (1987). *The sounds of French: An introduction*. Cambridge university press.

Tranel, B. (1995). Current issues in French phonology: Liaison and position theories. *The handbook of phonological theory*, pages 798–816.

Trask, R. L. (1996). *Historical linguistics*. Oxford University Press.

Trubetzkoy, N. (1939). *Grundzüge der Phonologie*. Prague, Vandenhoeck and Ruprecht [Translation: C. A. M. Baltaxe (1969). *Principles of Phonology*, Berkeley: University of California Press].

Truckenbrodt, H. (2007). The syntax-phonology interface. *The Cambridge handbook of phonology*, 435456.

Vendler, Z. (1968). *Adjectives and nominalizations*. Walter De Gruyter Incorporated.

Vogel, I. and Kenesei, I. (1990). Syntax and semantics in phonology. *The phonology-syntax connection*, 339:364.

Watson, J. C. (2011). *Word stress in Arabic*. Wiley-Blackwell.

Woliński, M., Miłkowski, M., Ogrodniczuk, M., Przepiórkowski, A., and Szałkiewicz, □. (2012). PoliMorf: a (not so) new open morphological dictionary for Polish. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012*, pages 860–864, Istanbul, Turkey. European Language Resources Association (ELRA).

Wright, S. K., Hay, J., and Bent, T. (2005). *Ladies first? Phonology, frequency, and the naming conspiracy*. Walter de Gruyter.

Wróblewska, A. (2018). Extended and enhanced Polish dependency bank in Universal Dependencies format. In de Marneffe, M.-C., Lynn, T., and Schuster, S., editors, *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 173–182. Association for Computational Linguistics.

Zec, D. and Inkelas, S. (1990). Prosodically constrained syntax. *The phonology-syntax connection*, pages 365–378.

Zwicky, A. M. and Pullum, G. K. (1986). *The principle of phonology-free syntax: introductory remarks*. Ohio State University. Department of Linguistics.