



KATHOLISCHE UNIVERSITÄT
EICHSTÄTT–INGOLSTADT

Dissertation

by Thomas Michael Rußer-Strobel

**Integrating Syntactic Covariates into Topic Modeling –
Using Universal Dependencies for Improved High-Level Summaries
of Customer Reviews**

First Supervisor:

Prof. Dr. Joachim Büschken

Second Supervisor:

Prof. Dr. Katja Gelbrich

Catholic University of Eichstätt-Ingolstadt, Germany
Ingolstadt School of Management (WFI)
Chair of BA, Sales Management and Marketing

Defense: 10.12.2025

*“To construct a sentence is to breathe life into an amorphous mass of words,
establishing a set of connections between them.”*

Tesnière (1959)

Contents

Contents	i
List of Figures	iv
List of Tables	vi
List of Abbreviations	vii
1 Introduction	1
1.1 Motivation	1
1.1.1 Different perspectives on text data	2
1.1.2 Text data analysis in Marketing	4
1.1.3 Combining “how” and “what” for Marketing Insight	5
1.2 Structure of the thesis	11
2 Two perspectives on text data “how” & “what”	12
2.1 “How” – syntactic structure of language	12
2.1.1 General motivation for considering structural elements of language . . .	12
2.1.2 Constituency and dependency formalism	13
2.1.3 Universal Dependencies	20
2.1.3.1 Historical development	21
2.1.3.2 Annotation principles	24
2.1.3.3 Usefulness of UD for topic modeling applications	30
2.2 “What” – topic modeling with Latent Dirichlet Allocation	33

2.2.1	Idea	33
2.2.2	Model definition	34
2.2.3	Model estimation	37
2.2.4	Model insights	40
3	Model development	44
3.1	Revisiting LDA's BoW within documents assumption	45
3.2	Relaxing the BoW assumption	47
3.2.1	Observed sequence of words	48
3.2.2	Observed sentence boundaries	50
3.2.3	(Quasi-) observed grammatical features	53
3.2.4	Summary	59
3.3	Labeled Universal Dependencies LDA	60
4	Empirical analysis	70
4.1	Data	70
4.2	Topic inference	78
4.2.1	Benchmark models	79
4.2.2	Model fit	80
4.2.3	Face validity of inferred topics	82
4.2.4	Topic uniqueness	86
4.2.5	Topic coherence	90
4.3	Explaining topic propagation through syntactic structure	94
4.3.1	Empirical effects of dependency relation types	95
4.3.2	Dependency-based visualization of topic propagation	98
4.3.3	Topic propagation patterns across datasets	102
5	Discussion	110
5.1	Main contributions	112
5.2	Limitations and future research directions	114
5.3	Conclusion	116

A	Universal Dependencies framework	119
A.1	Overview of UD relation types	119
A.2	Critique of the Universal Dependencies framework	121
B	MCMC for the UD-LDA model	125
B.1	Draw of ϕ	126
B.2	Draw of θ	126
B.3	Draw of β	127
B.4	Draw of z_n and ζ_n	127
C	Measuring topic coherence	130
	Appendices	118
	Bibliography	134

List of Figures

1.1	Positioning of this thesis.	2
1.2	Comparison of an American Restaurant review represented as a bag-of-words and as Universal Dependency trees.	8
2.1	An example of a constituency parse tree.	14
2.2	Dependency parse tree annotated with Universal Dependencies.	17
2.3	A Universal Dependencies (UD) parse of the sentence “We went to New York and had a great time” represented as an arc diagram.	24
2.4	An annotated UD parse tree illustrating the distinction between core structural types.	25
2.5	Directed Acyclic Graph of the LDA model.	36
2.6	Visualization of five documents represented as topic distributions within the topic simplex for the case of $T = 3$ topics.	42
3.1	Illustration of sentence structure representations and the potential paths for topic propagation.	62
3.2	Labeled Universal Dependencies LDA (labeled UD-LDA).	64
4.1	Text data pre-processing for the labeled UD-LDA.	71
4.2	Distributions of head-dependent distance across datasets.	74
4.3	Head-dependent distances in parsed sentences.	75
4.4	Share of most common UD types across datasets.	77

4.5	Distribution of topic overlap given model and dataset ($N = 20$, excluding stop-words).	89
4.6	Distribution of w2v coherence given model and dataset ($N = 20$, excluding stopwords).	94
4.7	Posterior distribution of β -coefficients in Luxury Hotel reviews dataset.	97
4.8	Posterior distribution of β -coefficients in Camping Tents reviews dataset.	97
4.9	Color-coded example reviews from the Luxury Hotels dataset.	99
4.10	Color-coded example reviews from the Camping Tents dataset.	99
4.11	Dependency-based visualization of topic propagation in sentences from the Luxury Hotels dataset.	101
4.12	Dependency-based visualization of topic propagation in sentences from the Camping Tents dataset.	101
4.13	Relation-level topic propagation frequencies in the Luxury Hotels and Camping Tents datasets.	104
4.14	Visualization of relation-specific topic propagation in a UD-annotated sentence.	108

List of Tables

1.1	Overview of Objectives and Datasets in Text-Based Marketing Research.	6
2.1	Typology of 37 basic Universal Dependency relations (UD Version 2).	26
2.2	Description of UD relations in an example sentence.	27
2.3	Three exemplary topics represented by their ten most likely terms.	43
3.1	Overview of topic modeling approaches incorporating (observed) structural features of text data.	58
4.1	Summary statistics of the three customer review datasets	74
4.2	Model fit as the log marginal likelihood of the data (per word).	81
4.3	Top 20 words in labeled UD-LDA on Luxury Hotels dataset	84
4.4	Top 20 words in classical LDA on Luxury Hotels dataset	84
4.5	Model level topic overlap given model and dataset, based on $N = 20$ most likely terms per topic (excluding stopwords).	88
4.6	Mean model-level Word2Vec coherence (Skipgram Window Size $\pm 4L$), based on $N = 20$ most likely terms per topic (excluding stopwords).	93
4.7	Pearson correlation coefficients of relation-level topic propagation frequencies across the three datasets.	103
4.8	Colorcoded relation-level topic propagation frequencies across all three datasets	106
A.1	Descriptions and examples of UD relation types.	119
C.1	Description of commonly used coherence measures.	133

List of Abbreviations

BoW	bag-of-words
CFG	Context-Free Grammar
DAG	Directed Acyclic Graph
EWT	English Web Treebank
LDA	Latent Dirichlet Allocation
LSI	Latent Semantic Indexing
NLP	Natural Language Processing
PLSI	Probabilistic Latent Semantic Indexing
POS	Parts-of-Speech
PSG	Phrase Structure Grammar
MCMC	Markov Chain Monte Carlo
UD	Universal Dependencies

Chapter 1

Introduction

1.1 Motivation

The value of text data in generating actionable insights is well recognized in marketing research (Berger et al. 2020). A central challenge in analyzing text data lies in its inherently interdisciplinary nature. On one hand, text can be viewed as a linguistic object, placing its analysis within the domain of language research or linguistics. On the other hand, texts represent a form of unstructured data, which can be analyzed statistically to uncover latent patterns and infer their semantic content – an approach situated within applied statistics or automated text processing. To derive marketing-relevant insights from text data, it is necessary to integrate these disciplinary perspectives into a unified analytical framework (Berger et al. 2020). Figure 1.1 illustrates the conceptual foundations combined in this dissertation to position it within the field of “Natural Language Processing (NLP) in Marketing.” At its core, this work contributes to marketing research, which is concerned with the systematic collection and analysis of data to draw inference about customers and market-level behavior (Lee & Bradlow 2011). Marketing analyses are inherently targeted, focusing on selected markets, customers, or product categories. To support such analyses, this thesis draws on methods from automated text processing – specifically, topic modeling. The application of topic models in marketing contexts has already demonstrated substantial promise (Reisenbichler & Reutterer 2019). In addition, this work incorporates methods and theories from language research, particularly syntactic parsing using the Universal

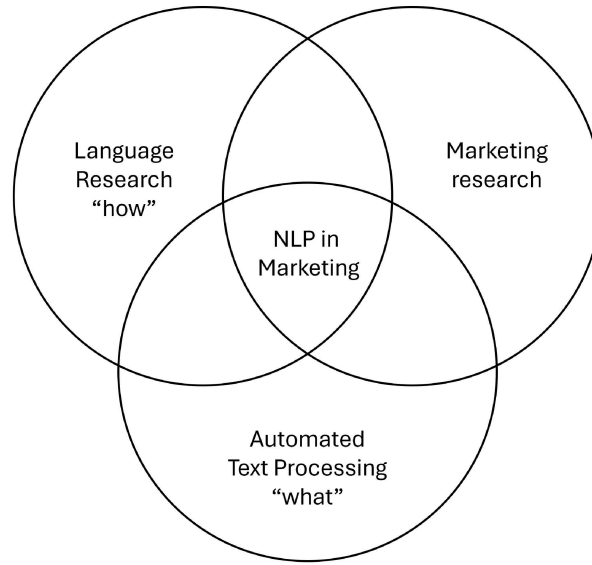


Figure 1.1: Positioning of this thesis at the intersection of Marketing Research, Language Research, and Automated Text Processing.

Dependencies (UD) framework (De Marneffe et al. 2021). The analysis of syntactic structures has also proven valuable in marketing-related studies (e.g., Atalay et al. 2023).

Syntactic structure is understood as a complementary layer of information that extends beyond surface-level word co-occurrence and can enhance topic inference. The integration of UD with topic models has not yet been systematically explored. This thesis addresses this gap by proposing **labeled UD-LDA**, an autocorrelated topic model conditioned on UD structures, and evaluating it against benchmark models using standard topic modeling metrics across three customer review datasets. It consistently achieved better model fit and more distinctive topics while maintaining competitive coherence, showing that syntactic constraints can enhance performance without sacrificing interpretability. Further analysis revealed that the type of syntactic relation systematically influenced topic propagation between words, providing a linguistic explanation for how the model captures thematic interdependencies within sentences.

1.1.1 Different perspectives on text data

In the context of this research, text data offers two complementary avenues for generating insights. The first focuses on identifying the specific content conveyed in texts – inferring “*what*”

underlying themes or topics are presented. The second examines the structural composition of texts – understanding “*how*” language is constructed and used (see Figure 1.1).

An approach primarily concerned with the “*what*” dimension of text falls under the domain of text mining or automated text processing. Netzer et al. (2012) define text mining as “*the process of extracting useful, meaningful, and nontrivial information from unstructured text.*” The central aim is to uncover informative content by detecting patterns in textual data. Simple techniques include counting word frequencies, for example, tracking how often a brand is mentioned in social media posts. More advanced methods, such as embedding models, represent words as multidimensional vectors in a continuous space, placing semantically related terms in close proximity based on contextual similarity (see, e.g., Mikolov et al. 2013). Topic models such as those introduced by Blei et al. (2003) and adapted for marketing by Büschken & Allenby (2016, 2020, 2025), identify latent themes by analyzing word co-occurrence patterns across documents.

A common characteristic of these approaches is extensive pre-processing of raw text. This may include removing rare words, stemming or lemmatizing tokens, or substituting synonyms. As a result, the analyzed text often differs significantly from its original form. However, the goal is not to examine linguistic structure per se, but to use text as a medium for uncovering underlying content-related phenomena.

In contrast, analytical interest may also center on “*how*” texts are structured. One of the defining features of text data is the flexibility authors have in expressing information: They may write in complete sentences or bullet points, adopt an emotional or neutral tone, and vary in syntactic complexity. When the primary focus lies in such structural aspects, text is no longer treated merely as unstructured data – it is understood as language. Accordingly, such methods are conceptualized in this study as *linguistic approaches* to text analysis.

Examples include Parts-of-Speech (POS) tagging, which identifies the grammatical categories of words in a sentence (e.g., Petrov et al. 2011, Kronrod et al. 2022); sentiment analysis, which quantifies the emotional tone conveyed in text (e.g., Rocklage et al. 2023); and syntactic parsing, which reveals the grammatical relationships among words in a sentence (e.g., Nivre

et al. 2020, De Marneffe et al. 2021). This illustrates how linguistic approaches examine the structure and usage of language.

It is important to underscore that language and text data can be approached from multiple, complementary perspectives. This does not mean that analytical methods belong strictly to one category or the other. Rather, these perspectives are often interdependent: The form (“*how*”) of linguistic expression influences its meaning (“*what*”), and vice versa. This interdependence underpins the rationale for integrating both perspectives when analyzing text in marketing contexts.

1.1.2 Text data analysis in Marketing

The analysis of text data in marketing is typically driven by a specific goal, which we refer to as a “*targeted analysis*”. This means that the examination of textual data is aligned with a clearly defined, marketing-relevant objective (see Table 1.1). For example, one objective might be to investigate how textual complexity (i.e., *how* language is structured) influences social media engagement (Pancer et al. 2019), or to assess the impact of syntax and sentence construction on the effectiveness of marketing messages (Atalay et al. 2023). Other objectives may center on content-based insights (i.e., *what* is being communicated), such as deriving perceptual brand maps from textual data (Netzer et al. 2012), or identifying latent topics in customer reviews to predict satisfaction ratings (Büschken & Allenby 2016). In all cases, marketing research begins with a clearly defined objective that shapes the subsequent text analysis.

The targeted nature of such analyses also affects the type and scope of data deemed suitable. Consider the following example: Suppose we operate an upscale steakhouse in Los Angeles, California, and our objective is to identify latent themes in customer reviews that contribute to satisfaction or dissatisfaction. In this context, analyzing a large and diverse collection of restaurant reviews from around the world would provide limited actionable insight. The reasons for (dis)satisfaction among patrons of various restaurant types, price points, and locations are too heterogeneous to support specific, context-sensitive decisions. Instead, the focus must lie in uncovering drivers of (dis)satisfaction among a defined target group – namely, guests of upscale steakhouses in the Los Angeles area. This clientele may have particular expectations, such as

the origin or aging process of the meat, which are only revealed through a **carefully curated and context-specific dataset**. In such a case, the dataset might include reviews restricted to a specific cuisine, price tier, and geographic region.

This pattern of deliberate data selection is also evident in prior marketing studies using text data (see Table 1.1). Datasets are often focused on narrowly defined product categories such as digital cameras (Lee & Bradlow 2011), sedan cars (Netzer et al. 2012), books (Ludwig et al. 2013), or oral care products (Timoshenko & Hauser 2019). Typically, these datasets consist of a few thousand documents – ranging from product reviews and social media comments to blog posts, essays, or articles. While these volumes exceed the scale of manual analysis, they remain modest compared to the massive datasets used to train general-purpose language models.¹ In short, the targeted nature of marketing-related text analysis inherently constrains the volume of usable data. This makes it all the more essential to use available textual information as efficiently as possible – an idea that will be further elaborated in the next section.

1.1.3 Combining “how” and “what” for Marketing Insight

The aim of this thesis is to identify latent topics – i.e., *what* is conveyed in customer reviews. To this end, probabilistic topic models such as Latent Dirichlet Allocation (LDA; Blei et al. 2003) are commonly employed. In LDA, topics are modeled as probability distributions over a fixed vocabulary, while each document (e.g., a customer review) is represented as a probability distribution over those topics. A foundational assumption of Latent Dirichlet Allocation (LDA) is the *bag-of-words (BoW)* assumption, which posits that the topic assignment of each word is conditionally independent of all others, given the document-level topic distribution. As a result, word order and syntactic structure are ignored. Topic inference relies solely on how often words appear together within a document, regardless of their position or grammatical relationship. Under this assumption, any ordering of words is admissible, as illustrated in Figure 1.2 (1) where the document is represented as a word cloud.

However, a comparison with the original review reveals that much of the information about *how* the text is structured is lost. The specific word order (e.g., readability), the presence or

¹For comparison, Brown (2020) introduce GPT-3, a 175-billion-parameter model trained on a dataset comprising 300 billion tokens.

Author(s)	Marketing Related Objective	Data Set(s)
Ludwig et al. (2013)	Investigate the influence of emotions and writing style in consumer reviews on online retailers' conversion rates	18,682 book reviews obtained from Amazon
Schweidel & Moe (2014)	Infer brand sentiment from posts conditional on the social media venue	7,565 comments related to an enterprise software brand; 4,173 comments related to a telco brand
Pancer et al. (2019)	Relation of textual complexity (readability) and social media engagement	4,000 posts from a single Facebook page in a two year period
Labrecque et al. (2020)	Investigate the relationship between linguistic (pronoun) choices and consumer engagement (liking, commenting and sharing) with social media posts	15,788 postings from the global top-100 brands in the year 2015
Atalay et al. (2023)	Investigate the role of syntax in designing effective marketing messages	13.7 million reviews of 500,000 different products available on Amazon; 1,012 transcribed conversations from the persuasion for good dataset
Lee & Bradlow (2011)	Describe market structure by inferring brand-attribute associations from phrases	8,226 digital camera reviews in a four year period
Netzer et al. (2012)	Infer market structure brand perceptual maps from UGC	568,174 sedan car forum messages; 670,000 diabetics drug forum messages
Büschken & Allenby (2016)	Infer topics from UGC that allow for predicting customer ratings	696 Italian restaurant reviews; 3,212 upscale hotel reviews; 1,255 midscale hotel reviews
Timoshenko & Hauser (2019)	Infer customer needs from UGC	115,000 oral-care reviews; 7,000 kitchen application reviews; 11,000 skin treatment reviews; 500,000 sentences from social listening about prepared food
Hong & Hoban (2022)	Predict each sentence's contribution to the success of creative appeals	60,943 essays from a crowdfunding platform for teachers

Table 1.1: Overview of Objectives and Datasets in Text-Based Marketing Research.

absence of punctuation, sentence structure, and grammatical relationships between words are all disregarded under the BoW assumption. In LDA, the only information considered relevant is co-occurrence of words at the document level.

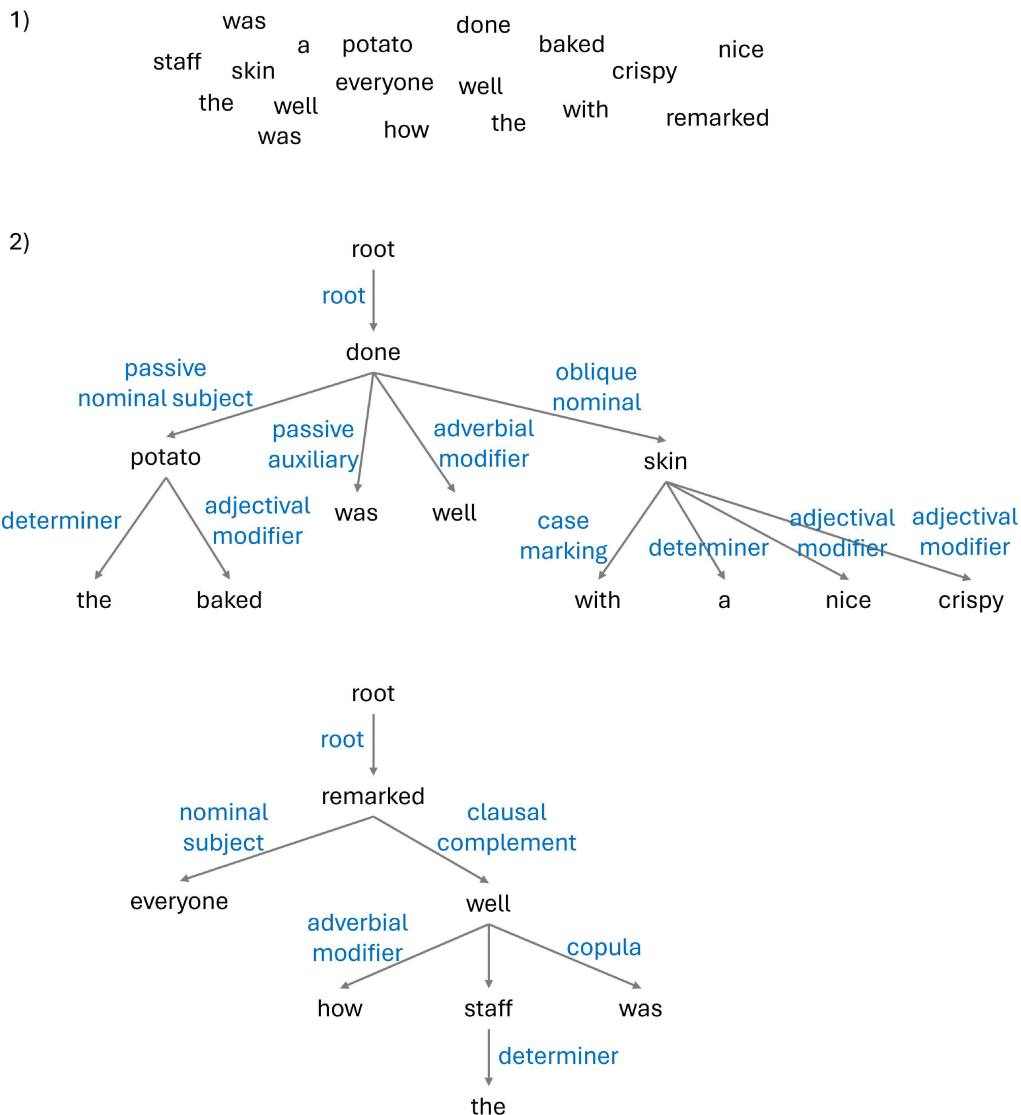
This thesis argues that such a limited perspective represents an inefficient use of available textual information. Text data inherently contain rich structural cues that extend beyond simple co-occurrence patterns. Incorporating such information into topic models can enhance topic inference by considering not only *what* is said, but also *how* it is expressed. Indeed, prior research has demonstrated the benefits of integrating structural features into topic modeling frameworks. For example, Wallach (2006) augment LDA with bigram information to model sequential word dependencies. Büschken & Allenby (2016) assign a single topic per sentence, using sentence boundaries as structural units. Büschken & Allenby (2020) model serial transitions in topics based on punctuation and word order. Other work has incorporated syntactic structure more explicitly: Balikas et al. (2016) use noun phrases and sentence boundaries to constrain topic assignment, while Boyd-Graber & Blei (2008) model topic transitions using unlabeled syntactic dependencies. These examples suggest that the structural dimension of text (i.e., the *how*) can be effectively employed to inform and refine topic inference.

Some structural features are directly observable (e.g., punctuation, sentence segmentation), while others require pre-processing, such as syntactic parsing. For the latter, standard software packages can be used during the data preparation phase. Figure 1.2 (2) presents an example of a syntactic dependency tree constructed using the Universal Dependencies (UD) framework (Nivre et al. 2020). UD is a cross-linguistic annotation framework that represents grammatical structure through labeled head-dependent relations between words. Its design facilitates consistent syntactic analysis across languages, providing a linguistically principled representation of sentence structure (see Chapter 2.1.3 for a detailed overview of the framework). Even at a glance, it is evident that the UD representation retains considerably more structural detail than the BoW format. Specifically, UD encodes two forms of structural information that are of particular interest for this study.

First, UD encodes grammatical relations as **directed dependencies between head and dependent words**. As shown in Figure 1.2 (2), these relations are not aligned with word order:

Figure 1.2: Comparison of an American Restaurant review represented as a bag-of-words and as Universal Dependency trees. The “cloud” representation in 1) indicates that any word order is admissible.

The baked potato was well done with a nice crispy skin.
Everyone remarked how well the staff was.



Notes. Top: Original Review. 1: Review in bag-of-words representation. 2: Review in Universal Dependency representation. Arrows indicate syntactic head-dependent relationships between words. Blue labels specify the syntactic type of head-dependent relationship.

they can span non-adjacent words and operate in either direction. For example, the dependency *potato* $\rightarrow$ *baked* runs counter to the reading direction, while *skin* $\rightarrow$ *nice* bypasses the intervening word *crispy*. These structures reveal words’ relationships which cannot be captured by bag-of-words models or models of sequential word order. Further, syntactic dependencies often reflect intuitive thematic relationships between words. In the relation *potato* $\rightarrow$ *baked*, the head word “potato” conveys the main topical content, while “baked” refines its meaning without introducing a separate topic. Similarly, the dependents “nice” and “crispy” refine the meaning of their head, “skin”.

This observation motivates a shift in how topic assignment is conceptualized. This thesis hypothesizes that topics are not independently assigned to each word, but may instead be *propagated* along the edges of syntactic dependency trees. This leads to the first research question:

RQ1: Does a topic model that conditions topic propagation on UD-based syntactic dependencies yield an enhanced representation of the latent topic structure in text datasets?

Second, Universal Dependencies are **typed** – each syntactic relation is annotated with a **label that specifies the type of dependency between head and dependent** (see Figure 1.2 (2)). These labels introduce an additional layer of linguistic information that may support more nuanced modeling of topic structure in text.

It is plausible that certain types of syntactic relations signal stronger thematic connections than others. For example, the UD framework defines various modifier relations such as *amod* and *advmod*, in which the dependent word refines or specifies the meaning of the head. In the relation *amod*(*skin*, *crispy*), the adjectival modifier “crispy” elaborates on the meaning of “skin”, indicating a close topical association.

By contrast, the relation *ccomp*(*remarked*, *well*) reflects looser thematic ties. In this case, “remarked” is the predicate in the main clause, while the dependent “well” belongs to a subordinate clause that functions as a clausal complement (“how well the staff was”). Crucially, the main clause (“everyone remarked”) remains grammatically complete and intelligible even

without the complement, which provides optional elaboration rather than essential information. This syntactic configuration implies a weaker dependency at the topical level, as the thematic continuity between head and dependent is less tightly coupled.

These examples support the hypothesis that the probability of topic propagation between syntactically linked words may vary depending on the type of dependency relation. This motivates the second research question:

RQ2: Does the type of syntactic dependency, as indicated by UD labels, contribute to explaining the probability of topic propagation along the edges of dependency trees?

To address these research questions, we develop a conceptual topic model that incorporates structural features of text (the *how*) into the inference of latent topics (the *what*). Specifically, we propose the **labeled UD-LDA**, a topic model that allows for topic propagation between syntactically connected words. While the mechanism of topic propagation builds on the autocorrelated structure introduced by Büschken & Allenby (2020), our approach introduces a key extension: we restrict topic propagation to head-dependent pairs identified through UD parsing. That is, topic propagation is only permitted along the branches of syntactic UD trees, and the probability of such topic propagation is conditioned on the type of dependency relation.

The labeled UD-LDA model is empirically applied to three customer review datasets representative of common use cases in marketing research. Its empirical performance is evaluated against classical LDA, which operates under the BoW assumption, and against models that incorporate alternative forms of structural information, such as word order and sentence boundaries.

The aim of this thesis is to explore the role of syntactic structures – specifically, labeled UD trees – in the modeling of topic dependencies. We seek to understand whether and how syntactic dependencies inform topic inference, and which relation types are most influential in driving or inhibiting topic propagation. This work does not aim to define the boundary conditions of the model, such as optimal document or sentence length for exploiting syntactic structure. Instead, it adopts an exploratory perspective to assess whether syntactic annotations provide meaningful signals for enhancing topic models.

1.2 Structure of the thesis

The structure of this study is as follows:

Chapter 2 approaches text data from two distinct perspectives. Section 2.1 adopts the “how” perspective, focusing on the syntactic and grammatical structure of language, with particular attention to the operationalization of language structure through the UD framework (Nivre et al. 2020). Section 2.2 then turns to the “what” perspective by introducing LDA (Blei et al. 2003), a widely used topic model for inferring latent themes in text.

Chapter 3 addresses the central tension between the BoW assumption in LDA and the syntactic structure of language captured by UD. Section 3.2 reviews topic model extensions that incorporate structural features of language into the topic inference process. Section 3.3 then presents the proposed integration of the “how” and “what” perspectives through the development of the *labeled UD-LDA*, a novel topic modeling approach that conditions topic inference on syntactic dependencies.

Chapter 4 presents an empirical evaluation of the labeled UD-LDA using three consumer review datasets representative of typical marketing applications. Section 4.2 assesses the model’s empirical performance against several benchmarks, focusing on model fit and topic quality. Section 4.3 investigates how different syntactic dependency types influence topic propagation at the sentence level.

Finally, **Chapter 5** concludes the thesis with a discussion of the findings on the relationship between linguistic form and content – and outlines key limitations and directions for future research.

Chapter 2

Two perspectives on text data “how” & “what”

2.1 “How” – syntactic structure of language

2.1.1 General motivation for considering structural elements of language

When people write about their experiences in consumer reviews, blog posts, discussion forums, or other media, they tend to formulate readable and understandable sentences. Tesnière (1959) observed that, unlike a dictionary, words within a sentence are not perceived in isolation.¹ Constructing sentences involves establishing connections between words which can be described in terms of syntax (grammatical structure) and semantics (meaning). Against this background, Futrell et al. (2020) define grammar as a *“function from form to meaning or from meaning to form, or in terms of an intermediate representation from which both meaning and form are derived”*. We view such a grammatical system as a bidirectional bridge between form and meaning, where meaning can be inferred from form and vice versa. Without a proper grammatical structure, understanding a sentence’s meaning becomes significantly more challenging (e.g. Tesnière 1959, Futrell et al. 2015, Atalay et al. 2023). This challenges the common assumption among data analysts – especially in bag-of-words models – that textual data are inherently unstructured. In reality, language follows specific syntactic patterns that shape both form and meaning. When

¹Note that an alphabetically ordered dictionary still exhibits greater structural organization than a “bag of words”.

learning a new language, a substantial part of the process involves studying its grammar. Humans rely on such grammatical knowledge – often applied implicitly in their native language – to interpret meaning. However, machines and algorithms lack this innate understanding unless it is explicitly embedded in the data modeling process. This underscores the need for a formally constructed, computationally processable framework that systematically represents the syntactic and semantic relationships within textual data.

2.1.2 Constituency and dependency formalism

A fundamental question concerns the effective representation of the inherent syntactic structure of sentences, especially for the purpose of using this type of information to improve topic modeling. **Constituency** (or phrase) structure and **dependencies** are two common representations of the linguistic structure of sentences. Phrase structure trees represent nested multi-word constituents, while dependency trees depict the relationships between individual words. We briefly describe both formalisms, highlighting some of their advantages and disadvantages, using the following example sentence that one can find in a typical consumer review of New York City hotel visitors.

we went to New York and had a great time

Constituency: Syntactic constituency holds that groups of words can function as single units, or constituents. To determine whether a sequence of words forms a constituent, various tests can be applied. For example, constituents can often be repositioned within a sentence, as in “*we [had a great time] and [went to New York]*”. They can also be substituted, as in: “*we [traveled] and had a great time*” (Jurafsky & Martin 2021). These tests highlight how constituency reveals the internal structure of sentences. A phrase is a specific type of constituent headed by a particular word class, and its type indicates the internal organization of the phrase. Among the most common are noun phrases (*NP*) and verb phrases (*VP*): noun phrases (*NP*) are made up of a sequence of words centered around at least one noun, exemplified by “a great **time**”. The verb phrases (*VP*) in English consist of a verb surrounded by various other elements, as illustrated by “**went** to New York”.

To identify the constituency structures of sentences, formal generative rules such as *Context-*

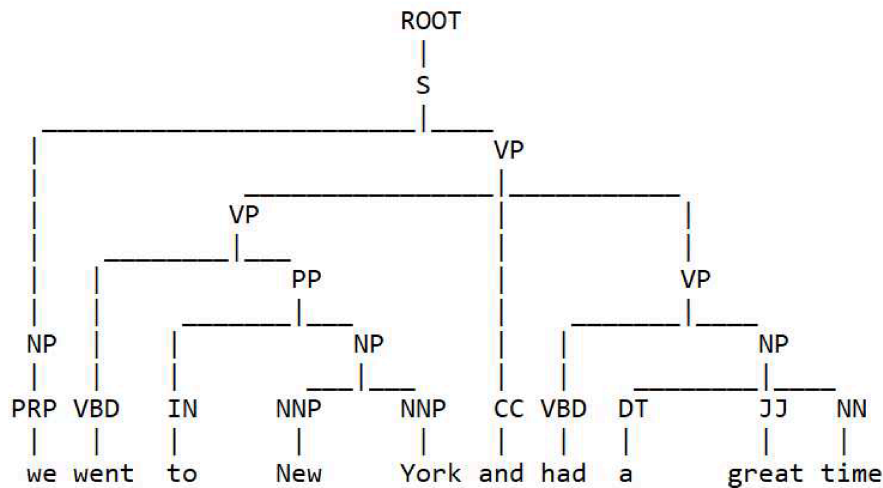


Figure 2.1: An example of a constituency parse tree.

Free Grammars (CFGs), also known as *Phrase Structure Grammars (PSGs)*, are required (Jurafsky & Martin 2021). A CFG serves both as a mechanism for generating sentences and as a tool to assign a grammatical structure to a given sentence (“constituency parsing”). A CFG consists of production rules ($A \rightarrow B$) that define valid symbol combinations. It distinguishes between non-terminal symbols, which represent grammatical categories (e.g., POS or phrases), and terminal symbols, which are the actual words of the language, typically listed in a lexicon. CFG rules allow for hierarchical embedding. For example, the rule $S \rightarrow NP VP$ states that a sentence (S) may consist of a Noun Phrase (NP) followed by a Verb Phrase (VP), as shown in Figure 2.1. A further rule like $NP \rightarrow NN|NNP|PRP$ specifies that a Noun Phrase (NP) may consist of a noun (NN), proper noun (NNP), or personal pronoun (PRP) - each representing a POS category. Introducing terminal rules, such as $NNP \rightarrow New|York$, connects the grammar to actual words, indicating that the proper noun (NNP) can be realized by words “New” and “York”. This hierarchical system supports the construction of tree-like structures to represent sentence grammar, as illustrated in Figure 2.1.

Collections of syntactically annotated sentences with constituency parses that have been reviewed and corrected by human experts are known as *treebanks*. The *Penn Treebank* (Marcus et al. 1993) is one of the most widely used resources for English. Treebanks serve as training data for publicly available parsing models, which are used to automatically generate constituency

parses from raw text. The sentence shown in Figure 2.1 was parsed using the *Stanza* Python library (Qi et al. 2020).

The analysis of the tree in figure 2.1 reveals a nested (recursive) arrangement of constituents. For example, the word “time” appears within a noun phrase (*NP*), which is embedded in a verb phrase (*VP*), itself nested within another verb phrase (*VP*), and ultimately within a sentence (*S*). As a result, each word may be associated with multiple nodes in the constituency tree, a property known as one-to-one-or-more correspondence. However, for certain applications, a flat, non-recursive structure that assigns each word to a single, clearly bounded unit may be more practical or computationally efficient. One method of avoiding the recursive nature of constituency trees is *partial parsing*. This strategy focuses on identifying non-overlapping, non-recursive elements within the text (Abney 1997). Rather than constructing a full recursive tree, it extracts only flat, non-nested constituents. This process can be facilitated by techniques such as *noun phrase chunking*, where information from the complete parse tree, as illustrated in Figure 2.1, helps identify noun phrase chunks. Specifically, sequences of Parts-of-Speech (POS) tags are used to define rules for chunking (e.g., Abney 1997, Bird et al. 2009). For example, in Figure 2.1, the sequence of a determiner (*DT*), followed by an adjective (*JJ*) and a noun (*NN*), signals a noun phrase chunk: [*a great time*]. Moreover, the non-terminal symbols located just above the POS tags in the full parse tree (such as *NP*) can further aid in identifying noun phrase chunks. The sentence, with noun phrases chunked and enclosed in brackets, would appear as follows:

[we] went to [New York] and had [a great time]

This non-recursive, flat representation is often more compatible with applications that require discrete, clearly bounded text units. For example, phrase-level groupings such as noun phrases can be treated as semantically coherent units, making them attractive for analyses that benefit from localized meaning (Balikas et al. 2016).

However, this simplification also comes with limitations. Important relational cues between words may be lost. For example, while the noun phrase “New York” might signal a reference to the city, its interpretation depends heavily on its connection to the verb “went”, which reveals

it as a destination for travel. Similar surface structures – like “we dined in New York” – could imply different types of activities (e.g., dining, lodging, sightseeing), depending on the verb. Such distinctions are preserved in the complete recursive constituency tree, where both the verb and the noun phrase are contained within a higher-level verb phrase (*VP*), such as “went to New York” (see Figure 2.1).

Another limitation of constituency-based trees, as in Figure 2.1, is that they do not explicitly represent the relationships between words within a phrase. Take, for example, the internal structure of the noun phrase (*NP*) “a great time”: All three words appear adjacent within the same constituent, but the tree provides no information about their relative importance. This raises the question of what “importance” means – whether it refers to the grammatical role that a word plays in defining the type of phrase or its semantic contribution to the phrase’s meaning. To address this, syntactic constituents can be linked to a lexical head using head-finding rules. The head is typically considered the grammatically most significant word in a phrase, that is, the word that determines the phrase’s syntactic category. For example, in the prepositional phrase (*PP*) “to New York”, traditional grammatical head-finding rules identify the preposition “to” as the head, as it determines the type of phrase (Miller & Miller 2011, pp. 49ff.). However, this approach is not without controversy. Johansson & Nugues (2007), for instance, assign head status to “New”, the first non-punctuation word following the preposition, placing greater emphasis on semantic content than grammatical structure. From this perspective, semantic heads are those words that contribute the most to the meaning of a phrase. In downstream applications that aim to infer *what* a text is about, identifying the semantically most informative word may be more useful, as it helps capture the meaning of a phrase. However, incorporating this semantic information into constituency parses requires additional pre-processing. Semantically motivated head-finding heuristics, such as those proposed by Johansson & Nugues (2007), provide a promising foundation. However, this approach also moves us closer to dependency structures, a framework explored in the following section.

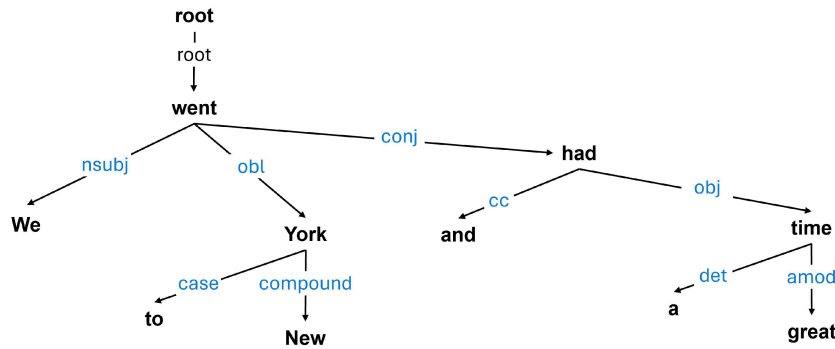


Figure 2.2: Dependency parse tree annotated with Universal Dependencies.

Dependency: A second family of formalisms used to represent sentence structure is dependency grammar. Contemporary theories of dependency grammar build largely on the foundational work of Tesnière (1959). The core principle of dependency grammars is that syntactic structure is defined through **directed, typed, binary relations** between words (Jurafsky & Martin 2021). The term **binary** means these dependencies link pairs of words only – dependencies between multi-word constituents are not defined, making words the primary unit of analysis. Dependencies are **directed**, with one word serving as the **head** and the other as the **dependent**. Moreover, dependencies are **typed**, meaning each dependency is annotated with a label that specifies the syntactic function of the dependent relative to the head. Consequently, the dependency structure of a sentence can be represented as a directed graph $G = (V, A)$, where V is the set of vertices corresponding to the words of the sentence, and A is the set of labeled, directed arcs representing dependencies (Jurafsky & Martin 2021). Figure 2.2 illustrates this concept with an example sentence represented as a dependency graph. A detailed description of abbreviations and examples is provided in Appendix A.1. These graphs, commonly referred to as *dependency trees*, exhibit several notable properties (Jurafsky & Martin 2021):

1. Each sentence contains exactly one root node – the only word word that does not depend on any other and instead functions as its own head.
2. Every other word in the sentence has precisely one head (i.e., one incoming arc), ensuring that all words are connected to the root via a unique path.

3. Each node V in the tree corresponds to exactly one word in the sentence, maintaining a strict one-to-one correspondence between words and nodes.

The structure of dependency trees – acyclic graphs with one head per word – reduces structural complexity compared to constituency trees, which involve nested constituents. This property is advantageous for text modeling, as dependency grammars describe grammatical relationships at the word level, matching the granularity used in such applications. For any word in a sentence, there is exactly one corresponding node in the dependency tree. The single incoming arc to this node indicates its head word and specifies the type of syntactic relationship. Consequently, each dependency can be represented as a labeled triple: $label(head, dependent)$ (De Marneffe & Manning 2008).

These dependency labels specify the syntactic relationship between a dependent and its head. The exact set of labels varies by dependency annotation scheme; for example, different schemes may use distinct levels of granularity or theoretical assumptions to represent syntactic phenomena. The examples in this thesis follow the **Universal Dependencies v2** annotation scheme, as described in De Marneffe et al. (2021). For instance, in Figure 2.2, the dependency from *York* → *New* is labeled as *compound*, indicating that the words “New” and “York” function morphosyntactically as a single concept. Similarly, the label *amod* on the dependency of *time* → *great* shows that “great” serves as an adjectival modifier of the noun “time” (De Marneffe et al. 2021). This type of information is particularly valuable in topic modeling because it provides insights not only into which words are connected in a sentence but also into the grammatical nature of these connections.

Another characteristic of dependencies is that they directly highlight **predicate-argument structures**, information that is not directly accessible in phrase structure parses (De Marneffe et al. 2006, Jurafsky & Martin 2021). An example is shown in Figure 2.2, where the sentence’s main predicate “went” is directly connected to its arguments “we” and “York”. In contrast, the constituency tree in Figure 2.1 does not make the connections between “we”, “went” and “York” immediately clear, as they span multiple layers. Predicate-argument structures help us to understand how the syntactic roles of words in a sentence link to meaning. As De Marneffe et al. (2021) explain, language reflects how humans perceive the world: we see entities (or objects) that

participate in events (actions and states). In language, the predicate (typically the head of a clause) expresses a state or action and can syntactically take arguments (typically nominals) representing the entities involved. In this example, “went” encodes the action described in the sentence. The relation *nsubj(went, we)* indicates that “we” is the nominal subject to the predicate “went”, while *obl(went, York)* identifies “York” as an oblique argument. This syntactic predicate-argument structure is not independent of the semantic interpretation of the sentence. De Marneffe et al. (2021) argue that surface-level syntactic predicate-argument relations correspond to underlying semantic roles: the syntactic subject often maps to the agent or actor in an event, while the oblique argument often corresponds to the patient or undergoer. In our example, “we” functions as the actor performing the action “went”, while “York” is the location being traveled to. Although predicate-argument structures describe syntactic roles, they often align closely with the semantic roles of the entities and events expressed in a sentence. Given this interplay between syntactic and semantic structure, the information encoded in dependency structures represents a promising resource for enhancing topic modeling.

Why Dependency? What we have examined so far serves as a high-level introduction to two distinct formalisms for capturing the syntactic structures of sentences. In both cases, our discussion has extended beyond purely grammatical aspects, touching on their connections to semantics. For constituency grammars, we noted that applying semantic head-finding rules can help highlight the semantically most significant word within multi-word constituents. For dependency grammars, we explored how syntactic predicate-argument structures map onto the semantic representation of events and the entities involved in them. It is therefore important to clarify the purpose of this thesis: the aim is not to describe sentence grammar in the most linguistically precise or accepted way, but rather to leverage syntactic structure to aid topic modeling – a task that is ultimately concerned with uncovering latent topics and thus leans more toward semantics than syntax. Nevertheless, both formalisms demonstrate that grammatical structure can serve as an indicator of how words in a sentence are connected not only syntactically but also semantically. Moreover, constituency and dependency grammars can, to some extent,

even be converted into one another (De Marneffe et al. 2006, Johansson & Nugues 2007, Futrell et al. 2020).

In the context of this thesis, we argue that the dependency formalism is better suited to our topic modeling use case. Dependency labels are particularly promising because they allow us to examine whether words are thematically closer or more loosely connected, depending on the type of syntactic dependency relation. Additionally, predicate-argument structures can directly link central words in a sentence, potentially helping to distill its core thematic content. Another factor supporting the focus on dependency formalisms is the prevailing trend in computational linguistics, which has increasingly moved from constituency to dependency treebanks (De Marneffe et al. 2021). From a technical perspective, the one-to-one correspondence inherent in dependency grammar aligns naturally with the word-level representation in topic models, potentially facilitating integration, even if this simplicity alone is not a decisive argument.

2.1.3 Universal Dependencies

In the context of dependency-based syntactic theories, this thesis draws on the Universal Dependencies (UD) framework. UD adopts a *lexicalist* perspective, treating the syntactic word as the fundamental unit of analysis. In written language, syntactic words do not always correspond to orthographic words separated by whitespace (Nivre et al. 2016). In UD, clitics or compounds, are segmented into individual syntactic words; for example, “I’m” is segmented into “I” and “m”. This emphasis on the word as the basic unit of analysis sets UD – and dependency frameworks more broadly – apart from constituency-based approaches.

UD is described as a framework for **morphosyntactic** annotation of human languages (De Marneffe et al. 2021), describing words by their morphological attributes and syntactic relationships. It defines two primary layers of annotation:

1. **Morphological layer:** documents the morphological properties of individual words, including the lemma, the parts-of-speech tag, and a set of features associated with the word form.

2. **Syntactic layer:** captures and labels a word’s dependencies to other words (head and dependent(s)).

The morphological layer specifies word-level properties such as the lemma (a word’s base form, typically found in dictionaries), POS tag (grammatical category), and features of the word form (e.g., number or tense). The syntactic layer, which is the primary focus of the UD guidelines and of this thesis, represents the dependencies between words, specifying their head-dependent connections and the type of syntactic relation. The goal is to leverage this dependency information to enhance topic modeling.

2.1.3.1 Historical development

To understand the UD Framework, it is essential to examine the six design criteria that guide its development. De Marneffe et al. (2021) describe these criteria as follows:

1. *“UD needs to be reasonably satisfactory on linguistic analysis grounds for individual languages — a journeyman’s universal grammar.*
2. *UD needs to be good for linguistic typology: It should bring out crosslinguistic parallelism across languages and language families.*
3. *UD must be suitable for rapid, consistent annotation by a human annotator.*
4. *UD must be easily comprehended and used by non-linguist users with prosaic needs.*
5. *UD must be suitable for computer parsing with high accuracy.*
6. *UD must support well downstream language understanding tasks, such as relation extraction, reading comprehension, machine translation, and so on.”*

These criteria highlight different dimensions of the value of a dependency grammar. The first two criteria focus on the framework’s suitability for linguistic analysis and typological research. Criteria 3 to 5 emphasize efficiency, aiming for ease of annotation, comprehensibility, and suitability for automated parsing. The final criterion addresses the framework’s utility for downstream analytical tasks. However, these criteria are not fully compatible and often involve

trade-offs. For example, achieving ease of understanding (Criterion 4) can come at the cost of linguistic precision (Criterion 1). As a result, not all of these (self-imposed) criteria can be equally well satisfied in practice.

The UD project emerged from multiple independent initiatives, notably the Stanford Dependencies annotation scheme (De Marneffe et al. 2006, De Marneffe & Manning 2008, De Marneffe et al. 2014), Google’s universal POS-Tags (Petrov et al. 2011), and the *interiset interlingua* for morphosyntactic tagsets (Zeman 2008).

Initial efforts focused on transforming existing constituency parses from treebanks, such as the Penn Treebank, into dependency parses. As a result, dependency structures and constituency structures are closely interconnected. Dependency representations offer a complementary perspective on sentence parsing by providing explicit details about predicate-argument structures (De Marneffe et al. 2006). According to De Marneffe et al. (2006), dependency structures are derived from phrase structures in two main steps: (1) *Dependency extraction*, which involves identifying the head of each constituent by choosing the semantic head over the syntactic head, and (2) *Dependency typing*, which labels the relationships.

The decision to designate content words as heads is a distinctive (and linguistically debated) feature of the Stanford Dependencies project, which was later carried over to the UD project. As noted by De Marneffe & Manning (2008), a key goal of Stanford Dependencies is to represent “semantically contentful relations”, aiming to improve their utility for downstream applications such as information extraction. Early in the project, the aim was also to create a straightforward description of grammatical relationships within sentences that could be easily understood by non-linguists. As a consequence, precise semantics were compromised for the sake of uniformity and simplicity (De Marneffe & Manning 2008), a decision that has attracted criticism from linguists, though it has been less controversial among practical users (see Appendix A.2).

The principles of uniformity and simplicity were also applied in related areas such as POS tagging. Zeman (2008) proposed a universal method for the automatic conversion of tagsets, introducing the “*interiset interlingua*” – an intermediate feature set designed to facilitate conversion from tagset A to tagset B. This approach laid the groundwork for converting treebanks with different annotation schemes into a unified representation. Petrov et al. (2011) introduced

a universal POS tagset comprising 12 categories intended to capture the most frequent POS distinctions across 22 languages. An early cross-linguistic effort in dependency annotation was undertaken by McDonald et al. (2013), who combined Stanford Dependencies (De Marneffe et al. 2006, De Marneffe & Manning 2008) with the Google universal POS tagset (Petrov et al. 2011) to develop a universal annotation scheme covering six languages. Building on these efforts, De Marneffe et al. (2014) propose an updated taxonomy of Stanford Dependencies, which had become the “de facto standard” in dependency representation. Their work focused on supporting cross-linguistic application and achieving a uniform representation, ultimately leading to the development of what they termed Universal Stanford Dependencies.

Nivre et al. (2016) describe UD (Version 1) as featuring cross-linguistically consistent annotation guidelines that had been applied to treebanks in 33 languages. Designed as an open community project, UD benefited from the participation of contributors who had previously worked on the Stanford Dependencies project and other cross-lingual annotation efforts. By the time Nivre et al. (2020) introduced UD (Version 2), treebanks were available for over 90 languages, supported by contributions from more than 300 researchers. De Marneffe et al. (2021) report that UD treebanks now cover more than 100 languages. Today, UD has become the cross-linguistic standard for dependency parsing (Jurafsky & Martin 2021).

The applicability of UD across different languages presents a compelling argument for their integration into topic modeling applications. Topic models are generally language-agnostic, treating texts as multinomial data (i.e., word counts), which allows them to be applied easily to texts in different languages. Integrating UD preserves this flexibility, allowing topic models to be used across languages without having to redefine the entire model for each one. This feature is particularly valuable for marketers: for example, multinational companies can analyze consumer reviews in different markets and languages without adapting the model specification. In the American market, they might process reviews in English, while in the German market, they would handle reviews in German. UD offer a unified framework that facilitates analysis in both contexts.²

²Unlike Universal Dependencies, there is no large-scale, cross-linguistic project for constituency grammar. Existing constituency resources are typically language-specific, and multilingual initiatives either involve converting UD annotations or aggregating heterogeneous treebanks (e.g., Li et al. 2025).

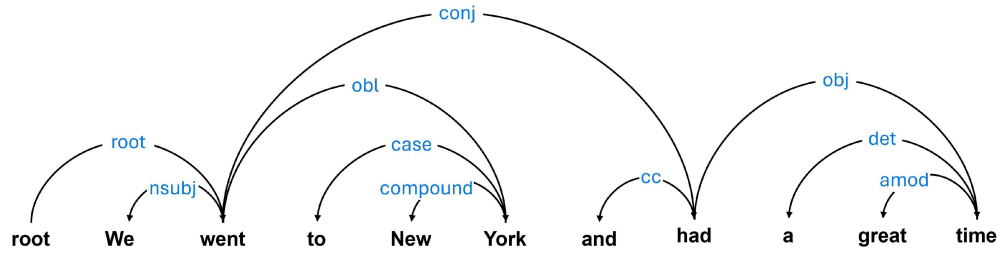


Figure 2.3: A UD parse of the sentence “We went to New York and had a great time” represented as an arc diagram.

2.1.3.2 Annotation principles

This thesis focuses on the syntactic layer of the UD framework, which encompasses grammatical relations between words. In this section, we will first discuss the syntactic information contained in UD. Secondly, we will examine how words are assigned the roles of head and dependent in pairwise relationships. Finally, we will outline the central characteristics of the identified dependencies.

Syntactic information The syntactic layer is based on the idea that humans perceive the world as consisting of entities (or objects) involved in events (actions or states), and that the syntactic structure of human languages reflects these semantic patterns (De Marneffe et al. 2021). To capture this, the UD syntactic layer distinguishes three basic types of structure:

1. **Clauses:** representing events or actions
2. **Nominals:** representing entities or objects
3. **Modifiers:** refining either nominals or clauses

Clauses express actions or states that typically involve participants realized as nominals, while modifiers can be used to refine both clauses and nominals. Together, these elements form a hierarchical structure in language. Beyond these core structures, UD also defines function words and additional syntactic elements.

Consider the sentence “*We went to New York and had a great time*” in Figure 2.3, annotated with Universal Dependencies.³ In its basic form, the dependency representation may not

³A detailed description of abbreviations and examples is provided in Appendix A.1.

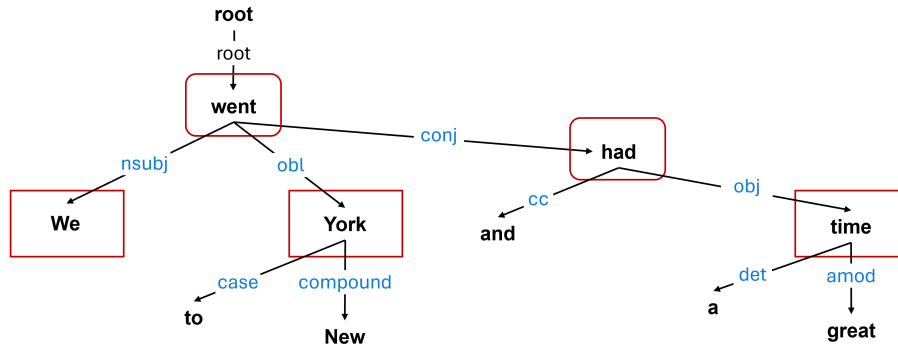


Figure 2.4: An annotated UD parse tree illustrating the distinction between core structural types. Rounded rectangles highlight the heads of clauses (predicates), while sharp-cornered rectangles indicate the heads of nominals (arguments).

immediately reveal how these three core structures are organized. Figure 2.4 is intended to guide the reader through the tree and provide an intuitive understanding of UD. Essentially, the sentence describes two events or actions – represented as clauses – with the head words “went” and “had” highlighted using rounded rectangles. Further, there are three entities involved in these events – represented as nominals – highlighted with sharp-cornered rectangles. The heads of these nominals are “we”, “York”, and “time”. The nominal “time” is modified by “great”, illustrating how modifiers refine entities. Similarly, modifiers can also refine clauses. The nominal “York” demonstrates a special syntactic relation to “New”, indicating that they form a compound of multiple words. Furthermore, function words such as “to”, “and”, and “a” attach to the head words of nominals and clauses. This illustrates how sentences are generally structured around clauses involving nominals, both refined by modifiers – an organization that aligns with the notion of syntactic predicate-argument structures De Marneffe et al. (2021).

These relationships among clauses, nominals, and modifiers are further refined in UD annotations through specific dependency labels. Table 2.1 organizes the 37 basic Universal Dependency relations (v2) according to whether the head word functions as the head of a clause (first two rows) or a nominal (third row), and whether the dependent is a nominal, clause, modifier, or function word (columns). The lower section includes special cases of relations that fall outside the narrow definition of dependency relations. The typology of these 37 basic grammatical relations can be further specified depending on the language to which it is applied, using a colon

to indicate subtypes. For example, in English, the relation *acl:relcl* specifies that *acl*, which generally denotes a clausal modifier of a noun, is further refined to indicate a relative clause.

	Nominals	Clauses	Modifier words	Function Words
Clausal Core arguments	nsubj obj iobj	csbj ccomp xcomp		
Clausal Non-core dependents	obl vocative expl dislocated	advcl	advmod discourse	aux cop mark
Nominal dependents	nmod appos nummod	acl	amod	det clf case
Coordination	Headless	Loose	Special	Other
conj cc	fixed flat	list parataxis	compound orphan goeswith reparandum	punct root dep

Table 2.1: Typology of 37 basic Universal Dependency relations (UD Version 2).

Adapted from: <https://universaldependencies.org/u/dep/> (accessed 2024-05-03).

To examine more closely the information conveyed by UD labels, let us again consider the example sentence shown in Figure 2.3. The dependency relation between the words “went” and “we” is labeled as *nsubj(went, we)*. This label, indicating a nominal subject, shows that “we” is the agent or doer of the event or action (De Marneffe et al. 2021). From Table 2.1, we can see that the dependent “we” is a nominal (i.e., an entity), while the head “went” has a clausal function (i.e., event or action). The nominal subject acts as a core argument of the clause it is connected to, further illustrating the concept of predicate-argument structures (see Chapter 2.1.2).

The object relation *obj(had, time)* indicates that “time” is a nominal and a core argument of “had”, which again serves as the head of a clause (De Marneffe et al. 2021). This reflects another predicate-argument relation, where the object “time” functions as the undergoer or patient of the action (i.e., the great time was had by someone).

The adjectival modifier relation *amod(time, great)* indicates that “great” is a modifier of the nominal “time” (De Marneffe et al. 2021). Here, “great” refines the meaning of “time”: they do

Dependency Relation	Description
amod(time, great)	“great” is the adjectival modifier of the nominal “time”.
case(York, to)	“to” is a (functional) case marking element to the nominal “York”.
cc(had, and)	“and” is a coordinating conjunction linked to its introduced conjunct “had”.
compound(York, New)	“New” and “York” are compounds that behave as a single unit.
conj(went, had)	“had” and “went” are two elements connected by a coordinating conjunction .
det(time, a)	“a” is a determiner (article) connected to the nominal “time”.
nsubj(went, we)	“we” is the nominal subject (core argument, typically a nominal) of the predicate “went”.
obj(had, time)	“time” is an object (core argument, typically a nominal) of the predicate “had”.
obl(went, York)	“York” is an oblique (non-core) nominal argument of the predicate “went”.
root(root, went)	“root” points to the (one!) root of the sentence: typically the sentence’s main predicate).

Table 2.2: Description of UD relations in the sentence: “we went to New York and had a great time”.

not just spend time in New York but have a great time. From these dependency relations, we can identify clear instances of all three basic UD structural types in our example sentence: nominals, clauses, and modifiers.

However, there are also relations that cannot be interpreted strictly as predicate-argument structures. One example is the determiner relation *det(time, a)*, which falls under the category of *function words* by attaching the determiner “a” to its nominal head “time”. As a function word, the determiner is subordinated to the nominal. This subordination of function words is a general principle in the UD framework: they are consistently treated as dependents in the relations in which they occur (De Marneffe et al. 2021).

Another special case is the relation *compound(York, New)*, which connects the two words “New” and “York” that morphosyntactically behave as a single unit, “New York”. Such multi-word constructions differ from typical modification (De Marneffe et al. 2021): “New” does not refine the meaning of its head “York” (as in distinguishing a new York from an old York). Instead, the combination of these two words creates a single, fixed meaning referring to the city “New York”. A summary and brief descriptions of all the ten UD relations identified in the sentence “we went to New York and had a great time” can be found in Table 2.2.

Head/Dependent assignments One important characteristic that has not yet been described concerns the question of which of the two words in a UD relation is superordinate or subordinate – that is, which word is designated the head and which the dependent. UD follow a clear guiding principle: content words are related through dependency relations, while function words attach to the content words they further specify (Nivre et al. 2016). According to De Marneffe et al. (2021), words in the main syntactic relations are generally assumed to be **autosemantic**, meaning they are content words with independent meaning – typically verbs, nouns, adjectives, or adverbs. This implies that content words tend to serve as heads in UD relations, while words of a more functional nature are treated as dependents. Essentially, this results in a notion of semantic hierarchy within dependency trees: words higher in the tree are more autosemantic than those lower in the tree (see Figure 2.4). For example, we see main dependencies connecting the semantic content words “we”, “went”, “York”, “had”, and “time”, while “York” and “time” are further specified by the words “New” and “great”. Function words such as “to”, “and”, and “a” carry limited meaning and appear at the lowest level of the dependency tree, not serving as heads of other dependents.

Although Universal Dependencies define syntactic relations via their labels, the framework’s principle of assigning autosemantic content words as heads and subordinating function words results in a hybrid of syntactic and semantic annotation. This design choice is motivated by the goal of increasing cross-linguistic parallelism, based on the assumption that languages are more alike in *what* is being said than in *how* it is being said (Croft et al. 2017, De Marneffe et al. 2021). However, the use of semantic heads is a major point of criticism among theoretical linguists, a critique that is addressed in Appendix A.2 of this thesis.⁴

Characteristics of identified dependencies: It should be noted that the first column of Table 2.2 highlights several relations that link words in predicate-argument structures, such as *nsubj(went, we)*, *obl(went, York)*, and *obj(had, time)*. These relations alone already convey the core content of the sentence. From them we can deduce that two distinct events are described

⁴A recurrent critique of UD concerns its decision to treat content words as heads over function words, a choice that blends semantic and syntactic criteria. Linguists argue this contradicts established syntactic principles and lacks historical precedent for subordinating elements such as prepositions and auxiliaries. While problematic for purely syntactic research, this structure creates direct links between semantically relevant words, which can be advantageous for certain downstream applications.

as well as the entities involved in each. The labels further specify the semantic roles of these entities: nominal subject “we” functions as the *agent* (or *doer*) in the “went” event, while the oblique object “York” serves as the *undergoer* in the same event; similarly, the object “time” functions as the *undergoer* in the “had” event.

Once these dependencies are established, the connected words are no longer interchangeable without altering the grammatical and semantic integrity of the sentence. For instance, in the alternative sentence “*As time went by, we had New York on our list to visit*”, the entity “time” participates in a different event and functions as the agent performing the action “went by”. This observation serves as an initial indication that word pairs involved in the main syntactic relations capture the gist of a sentence by forming its core predicate-argument structures. These structures identify which events are being described and which entities are actively or passively involved. Figure 2.4 illustrates this idea: the heads of clauses (used to express events) are directly connected to the heads of nominals and other clauses (used to express entities). These dependencies, which occur at the top of the tree, link the main content words of the sentence. All other relations either refine the meaning or ensure grammatical well-formedness.

We also observe that word pairs connected by dependency relations can vary considerably in their degree of semantic relatedness. At one extreme, the relation *compound*(York, New) effectively denotes a single semantic unit, since the two words jointly refer to the city “New York”. Other dependency pairs, such as *conj*(went, had), link elements that are semantically less tightly connected, representing two potentially independent events or actions. In contrast, certain relations such as *det*(time, a) reflect grammatical dependencies without clear standalone semantic contribution: the determiner “a” derives its function from syntactic necessity rather than semantic content. This observation, that different types of dependency relations encode varying degrees of semantic proximity between words, may be particularly useful for downstream applications such as topic modeling.

Dependency-linked word pairs are not constrained to adjacent positions in the surface word order (i.e., bigrams of consecutive words). This becomes especially evident when UD are visualized in linear reading direction, as shown in Figure 2.3. Dependency relations may span

both closely and distantly positioned words: for instance, “went” and “York” are separated by two intervening words, while “went” and “had” are separated by four. In contrast, some relations involve adjacent words, such as “went” and “we”, or “time” and “great”. Importantly, the direction of dependency relations does not necessarily align with the reading direction. For example, in *amod(time, great)* or *det(time, a)*, the dependency relation is directed against the reading direction. These observations suggest that grammatical dependencies are not adequately captured by models assuming first-order word dependencies in the reading direction. In contrast, the structure of UD enables the identification of linguistically meaningful word relations beyond simple heuristics, such as first-order dependencies in surface word order.

2.1.3.3 Usefulness of UD for topic modeling applications

This subsection summarizes the key arguments previously presented in support of using the UD framework, specifically in the context of its planned use for topic modeling.

As outlined in Section 2.1.2, dependency formalisms offer a number of advantages over constituency-based approaches, particularly in computational applications. Among these, the structural alignment between dependency grammars and topic models is especially notable. The one-to-one correspondence in dependency grammars, where each syntactic node corresponds to a single word, matches the granularity of topic models, which treat words as the fundamental unit of analysis. While topic models rely only on word counts at the document level, dependency structures offer two additional dimensions of information at the word level: (1) the identity of the syntactic head and (2) the type of grammatical relationship between a word and its head. This facilitates a direct and interpretable integration of syntactic information into topic models.

Another important advantage of the UD framework lies in its cross-linguistic applicability. Topic models are inherently language-agnostic and can be applied to text data in various languages without requiring changes to the underlying model architecture. The Universal Dependencies project supports this by providing treebanks in over 100 languages (De Marneffe et al. 2021). Consequently, topic models enhanced with UD annotations can be deployed in multilingual settings, an important consideration for international applications, such as the analysis of customer reviews across markets.

A further key feature of UD is its emphasis on predicate-argument structures, which reflect how humans conceptualize the world: as events (clauses) involving entities (nominals). The primary dependencies in a sentence, typically linking predicates to their core arguments, capture this semantic core effectively. Modifiers, in turn, attach to either clauses or nominals to enrich or specify meaning. This structuring of linguistic information is potentially valuable for identifying latent themes in textual data.

It is particularly noteworthy that grammatically related word pairs in a sentence are not necessarily adjacent. These dependencies do not follow simple heuristics such as fixed word windows or unidirectional relations aligned with reading order. In UD, syntactic relations can hold between neighboring and distant words, and the direction of the head-dependent relation may follow or oppose the reading direction. This characteristic opens up opportunities for enhancing topic models by moving beyond highly simplifying assumptions. Rather than modeling local thematic dependencies through proxies such as n-grams or sentence boundaries, UD allows for directly incorporating surface-syntactic word dependencies into topic models, thus providing a more linguistically grounded approach.

In addition, one of the most frequently voiced criticisms of UD from a theoretical linguistic standpoint also offers a valuable opportunity for integration with topic models: namely, the deliberate selection of content words as heads. In UD, head-dependent decisions are not based solely on grammatical criteria, but are also guided by the principle of *autosemantic properties*, the extent to which a word carries an independent semantic meaning (De Marneffe et al. 2021). Consequently, heads are generally more autosemantic than their dependents. This design choice introduces a semantic hierarchy into the syntactic tree structure, whereby words appearing higher in the tree possess more inherent semantic content, while words lower in the tree are more context-dependent.

This semantic argument is particularly compelling from a topic modeling perspective, as it enables modeling along this semantic hierarchy. Words with strong autosemantic properties can be assigned a topic in analogy to the i.i.d. topic assignment mechanism in LDA, where a topic assignment z is sampled from the document-level distribution θ_d . In contrast, less autosemantic words, those whose meaning is shaped by context, may inherit the topic of their syntactic head,

as indicated by the dependency structure. Thus, the semantic hierarchy encoded in UD trees offers a mechanism to condition the assignment of topics in the syntactic context.

Grammatical labels in UD further enrich this representation by specifying the syntactic nature of the head-dependent relationship. As Futrell et al. (2020) suggest, grammar can be conceptualized as “*a function from form to meaning or from meaning to form, or in terms of an intermediate representation from which both meaning and form are derived*”. In this sense, UD labels may indicate not only the syntactic function but also the semantic closeness between head and dependent. This opens up the possibility of leveraging grammatical relations to model topic dependencies at a more granular level, e.g., deciding whether head and dependent should share the same topic or have distinct topic assignments based on their syntactic link.

2.2 “What” – topic modeling with Latent Dirichlet Allocation

Having previously examined the syntactic structure of textual data – essentially addressing “*how*” text is constructed – this section shifts focus to analyzing “*what*” is expressed through text. Specifically, we explore the underlying thematic content conveyed in corpora of textual data. To this end, we take a closer look at Latent Dirichlet Allocation (LDA) which was introduced by Blei et al. (2003) in their article as the *smoothed LDA*, and hereafter referred to simply as LDA. LDA is arguably the most widely used topic modeling technique and serves as the foundation for numerous more advanced variants. Its primary use case is to uncover the latent thematic structure of text collections, such as consumer reviews (e.g., Büschken & Allenby 2016, 2020, 2025). However, LDA-based models have also been successfully applied to other types of discrete data for tasks such as image classification (e.g., Wang et al. 2009) or the analysis of supermarket shopping basket data (e.g., Hruschka 2014). LDA constitutes the basis for the topic model variant to be developed in this thesis. Therefore, a detailed examination of the LDA will be conducted, with a focus on textual data.

2.2.1 Idea

Targeted marketing analyses typically begin with a specific research objective in mind – be it understanding customer preferences, identifying unmet needs, or evaluating brand perception. The collection of textual data is carefully aligned with these goals. For example, researchers may collect customer reviews, social media posts, or survey responses related to a particular product category, brand, or consumption context. Once collected, the resulting body of text is referred to as a *corpus* which is divided into individual *documents*, such as customer reviews, each defined by observed textual boundaries. Documents, in turn, consist of sequences of *words*, which form the basic observed units in text analysis.

Beyond these observable elements, topic modeling introduces a layer of latent structure: it assumes that each document is a mixture of latent topics, and each topic is characterized by a specific probability distribution over words. For example, restaurant reviews might simultaneously address topics such as food quality, service, pricing, and atmosphere, although with varying emphasis. While some customers may focus primarily on the taste and presentation of

dishes, others might comment more extensively on the friendliness of the staff or the cleanliness of the venue.

From a marketing research perspective, the ability to discover such latent topics holds great analytical value. First, it offers a compressed, yet interpretable representation of large text corpora: What are the dominant themes that emerge across customer reviews? What topics are particularly relevant for certain customer segments? Second, it enables scalable analysis. Manually reading and annotating every review is impractical, if not infeasible. Topic modeling addresses this challenge by statistically inferring the underlying topics of the corpus and estimating the proportion of each topic within individual documents.

2.2.2 Model definition

LDA can be viewed as an extension of Latent Semantic Indexing (LSI) and Probabilistic Latent Semantic Indexing (PLSI). LSI (Deerwester et al. 1990) performs a *reduction in dimensionality* of textual data by decomposing the term document matrix ($V \times D$) into a term topic matrix ($V \times T$) and a topic document matrix ($T \times D$) using singular value decomposition (SVD). Here, V denotes the number of unique terms in the vocabulary, D the number of documents in the corpus, and T the number of latent topics. With typically $T \ll V$, this mathematical transformation yields a more compact representation of the text corpus. Building on this, PLSI Hofmann (1999) introduces a probabilistic perspective by modeling each word in a document as being generated from a document-specific *mixture of latent topics*. LDA as introduced by Blei et al. (2003) refines this idea by treating the topic mixture weights themselves as latent variables on the document level. This refinement is referred to as the *smoothed LDA*, a fully Bayesian *generative probabilistic model* in which Dirichlet priors are placed on both the distribution over topics for each document and the distribution over words for each topic. This smoothed version of LDA is the model most frequently cited in the literature and widely adopted in practical applications. It serves as a modeling technique for achieving a parsimonious and interpretable representation of a text corpus, one in which the generative process and its latent variables can be clearly specified and analyzed.

The generative process of LDA can be described as follows:

1. For each topic $t \in [1 : T]$:
 - (a) Draw ϕ_t from $\text{Dirichlet}(\beta)$ $\forall t$ i.i.d.
2. For each document $d \in [1 : D]$:
 - (a) Draw θ_d from $\text{Dirichlet}(\alpha)$ $\forall d$ i.i.d.
 - (b) For each word $w_{d,n}$ $n \in [1 : N_d]$:
 - i. Draw $z_{d,n}$ from $\text{Multinomial}(\theta_d)$
 - ii. Draw $w_{d,n}$ from $\text{Multinomial}(\phi_{t=z_{d,n}})$

where:

$w_{d,n}$ is the (observed) n -th word in document d

$z_{d,n}$ is the (latent) topic assignment of word n in document d

θ_d is the (latent) topic probability distribution for document d

ϕ_t is the (latent) word probability distribution for topic t

α and β are the (a priori fixed and symmetric) hyper-priors of θ_d and ϕ_t , respectively.

The fundamental idea of LDA is that each topic is characterized by a probability distribution ϕ_t over (unique) terms, that is, over the vocabulary. Additionally, each document is modeled as a random mixture of latent topics, denoted θ_d . This allows documents to contain varying proportions of a finite set of topics. As the term “latent” suggests, the topics themselves are not directly observed; the only observed variables are the words within documents. Each observed word $w_{d,n}$ is assumed to be generated through a two-step stochastic process: first, a topic indicator $z_{d,n}$ is sampled from the document-specific topic distribution θ_d ; second, given the topic assignment $z_{d,n}$, the word is sampled from the topic-specific word distribution $\phi_{t=z_{d,n}}$. This generative mechanism implies that different topics assign different probabilities to terms in the vocabulary. For instance, in the context of restaurant reviews, the topic “service” is more likely to

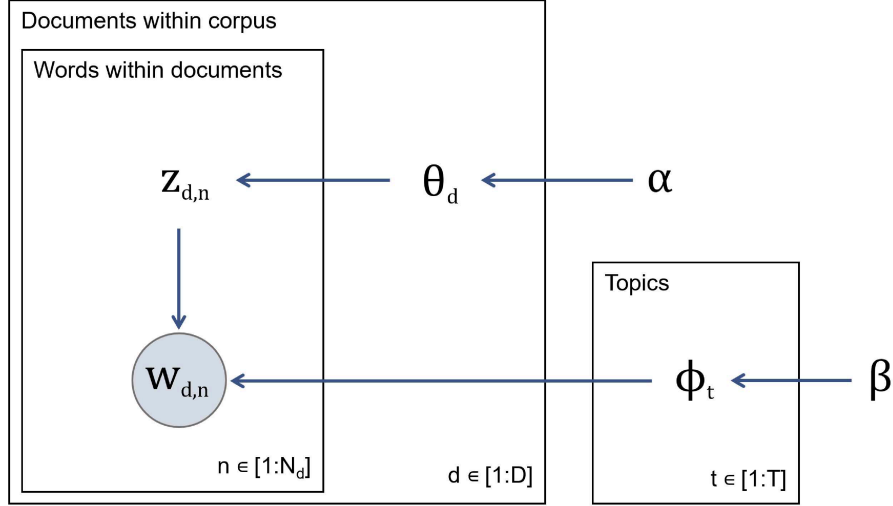


Figure 2.5: Directed Acyclic Graph of the LDA model.

generate words such as *waiter*, *friendly*, or *courteous*, whereas words like *delicious*, *crowded*, or *location* may be more probable under different topics such as “food” or “ambiance”. In this way, LDA explains the systematic co-occurrence of sets of words by attributing them to underlying latent topics.

The generative process of LDA can also be represented using a Directed Acyclic Graph (DAG), as shown in Figure 2.5. In this graphical model, the shaded node denotes the observed variables, namely the words in a document $w_{d,n}$. All latent (unobserved) variables are represented as unshaded nodes. The directed edges indicate (stochastic) dependencies between variables, while the plates denote replication over documents (d) and words within documents (n). Conceptually, the DAG is equivalent to the generative process described above. However, this representation is particularly useful for identifying conditional independence relationships among variables. For example, given the topic assignment $z_{d,n}$, the observed word $w_{d,n}$ is conditionally independent of the document-level topic distribution θ_d , since the relationship between θ_d and $w_{d,n}$ is fully mediated through $z_{d,n}$. The acyclic nature of the graph also facilitates an intuitive understanding of the hierarchical structure of the model. The directed edges clearly delineate the upstream (prior) and downstream (likelihood) components associated with each variable in the generative process.

Given the DAG of the LDA in figure 2.5 we can derive the factorization of the joint distribution of knowns and unknowns:

$$\begin{aligned}
 p(\phi, \theta, z, w, \alpha, \beta) &= p(\phi|\beta) \times p(\theta|\alpha) \times p(z|\theta) \times p(w|\phi, z) \times p(\alpha) \times p(\beta) \\
 &= \prod_{t=1}^T p(\phi_t|\beta) \left[\prod_{d=1}^D p(\theta_d|\alpha) \left(\prod_{n=1}^{N_d} p(z_{d,n}|\theta_d) p(w_{d,n}|\phi_{1:T}, z_{d,n}) \right) \right] p(\alpha) p(\beta) \quad (2.1)
 \end{aligned}$$

The hyperparameters α and β are typically fixed a priori and are commonly specified as symmetric *Dirichlet* priors. This choice reflects the assumption that, a priori, all topics are equally likely to be associated with any given document, and all words are equally likely to be associated with any topic. Although hyperparameters are not part of the inference process itself, they play an important role in shaping the model’s behavior through their smoothing effect (Heinrich 2008). Smaller values of α encourage the model to represent documents using a few dominant topics. In contrast, larger values of α will have a smoothing effect, resulting in a more uniform distribution of topics within documents. Similarly, lower values of β result in sparser word distributions within topics, making each topic more focused on a limited set of characteristic words. Higher values of β lead to more uniform topic-word distributions. However, since the optimization of these hyperparameters is not the focus of this thesis, further discussion will be omitted.

In summary, LDA allows for a parsimonious summary of textual data. Instead of characterizing documents by their raw word frequencies in a high-dimensional space of size V (vocabulary), LDA models documents in terms of a latent topic structure of size T (number of topics), where typically $T \ll V$ (Darling & Song 2013). In order to achieve this dimensionality reduction, the latent topic structure must be inferred from the observed text corpus.

2.2.3 Model estimation

LDA can be conceptualized as an unsupervised learning approach. The documents in the corpus are not associated with predefined labels, keywords, or categories. Instead, the algorithm infers the latent topic structure that is most likely to have generated the observed data. According to

Blei (2012), estimation techniques for LDA can broadly be divided into two categories: sampling-based methods and variational inference techniques. This section focuses on a sampling-based method for estimating LDA parameters, specifically (full) Gibbs sampling.

Model inference in LDA refers to reversing the generative process described above to estimate the posterior distributions of the latent model parameters θ , ϕ , and z , given the observed words w . The Dirichlet hyperparameters α and β are assumed to be fixed and known. The posterior distribution can be expressed as follows:

$$p(\theta, \phi, z | w, \alpha, \beta) = \frac{p(\theta, \phi, z, w | \alpha, \beta)}{p(w | \alpha, \beta)} \propto p(\theta, \phi, z, w | \alpha, \beta) \quad (2.2)$$

The denominator in the posterior expression is the marginal probability of the observed data (also referred to as *evidence*), which is intractable to compute (Blei et al. 2003, Darling 2011, Blei 2012). Consequently, the posterior distribution itself is also intractable to compute in a closed form.

However, the posterior distribution is proportional to the joint distribution of θ , ϕ , z , and w . Although the joint distribution does not belong to a standard family of distributions from which direct sampling is possible, it can be approximated using a Markov Chain Monte Carlo (MCMC) algorithm known as *Gibbs Sampling*. Gibbs Sampling allows to indirectly generate samples from a complex joint distribution without computing its full density (Gilks et al. 1995, Casella & George 1992). Instead of sampling from the joint distribution directly, one iteratively samples from the full conditional distributions of the latent variables z , θ , and ϕ . Each full conditional specifies the probability of a single latent variable, conditioned on all other variables and the observed data. This results in the following Gibbs sampling scheme:

$$p(z | w, \phi, \theta) \propto p(w | \phi, z) \times p(z | \theta) = \phi^{(v=w)} \circ \theta_d \quad (2.3)$$

$$p(\theta_d | z, \alpha) \propto p(z | \theta_d) \times p(\theta_d | \alpha) = \text{Dirichlet}(\alpha + C_{\cdot d}^{TD}) \quad \forall d \in \{1, \dots, D\} \quad (2.4)$$

$$p(\phi_t | w, z, \beta) \propto p(w | \phi_t, z) \times p(\phi_t | \beta) = \text{Dirichlet}(\beta + C_t^{WT}) \quad \forall t \in \{1, \dots, T\} \quad (2.5)$$

The order in which the full conditionals (2.3 to 2.5) are iterated is theoretically irrelevant. If

the process begins with sampling the topic assignments z , the two matrices ϕ and θ must first be (randomly) initialized.

The expression $\phi^{(v=w)} \circ \theta_d$ denotes the element-wise multiplication of the word distribution vector for word w with the topic proportions vector for document d , yielding a topic probability vector of length T . These probabilities serve as the parameters for a multinomial distribution, from which topic assignments z can be drawn independently for each word in each document, conditional on the current values of ϕ and θ .

Once the topic assignments z have been sampled, two count matrices can be updated. The topic-document count matrix C^{TD} records how often a topic $z = t$ is assigned to document d , while the word-topic count matrix C^{WT} tracks how often a word $v = w$ is assigned to topic $z = t$. These count matrices constitute the sufficient statistics for sampling from the full conditional posterior distributions of θ_d and ϕ . Given the topic assignments z , the draws of ϕ and θ are conditionally independent. Both are sampled from Dirichlet distributions whose parameters are obtained by adding observed counts to the prior pseudo-counts α and β , respectively. By repeatedly iterating through these full conditional distributions, a Gibbs chain of R samples is generated:

$$(\{z\}^{(0)}, \theta^{(0)}, \phi^{(0)}), (\{z\}^{(1)}, \theta^{(1)}, \phi^{(1)}), \dots, (\{z\}^{(i)}, \theta^{(i)}, \phi^{(i)}), \dots, (\{z\}^{(R)}, \theta^{(R)}, \phi^{(R)}) \quad (2.6)$$

The superscript i denotes the i -th iteration of the Gibbs sampler. Each draw of $\{z\}^{(i)}$ depends on the previous draw of $\theta^{(i-1)}, \phi^{(i-1)}$, while the draws of $\theta^{(i)}$ and $\phi^{(i)}$ are conditioned on $\{z\}^{(i)}$. This dependency structure induces the Markov property, whereby each state of the chain depends solely on its immediate predecessor, resulting in autocorrelation between successive samples. The early samples are still influenced by the (random) initialization, a phase commonly referred to as the burn-in period. If the number of iterations R is sufficiently large, the Markov chain converges to its stationary distribution, and the generated samples approximate draws from the joint posterior distribution of z , ϕ and θ (Casella & George 1992, Steyvers & Griffiths 2007).

Convergence can be assessed by monitoring the log-likelihood throughout the sampling process (Darling 2011).

A disadvantage of the full Gibbs model estimation procedure is the high autocorrelation in the Markov chain and the resulting computational inefficiency. For this reason, the collapsed Gibbs sampler is more commonly employed for LDA estimation. In this variant, the latent parameters ϕ and θ are analytically integrated (“collapsed”), which accelerates convergence. Collapsed Gibbs sampling for LDA is discussed in detail by Griffiths & Steyvers (2004), Heinrich (2008), and Darling (2011). Nevertheless, the full Gibbs approach remains valid and is particularly useful when working with extended or more complex variants of LDA for which collapsing ϕ and θ is no longer feasible.

Before turning to the interpretation of modeling results, it is worth briefly addressing the choice of the optimal number of topics T . Importantly, inference regarding T is not part of the Gibbs sampling procedure itself; rather, the number of topics must be specified in advance. Several strategies have been proposed to support the selection of an appropriate value for T . A common approach involves estimating multiple LDA models with different values of T and selecting the one that yields the highest log-likelihood on the training data (e.g., Griffiths & Steyvers 2004). Alternatively, T can be selected based on predictive performance on hold-out data, using metrics such as perplexity (e.g. Blei et al. 2003). A more advanced approach involves nonparametric topic models that infer the number of topics directly from the data (e.g., Yee Whye Teh & Blei 2006; Boyd-Graber & Blei 2008). However, identifying the optimal number of topics is not the primary objective of this work. Given the researcher’s familiarity with the datasets under investigation, the number of topics is fixed within a range considered appropriate for the research goals at hand.

2.2.4 Model insights

Overall model fit Model fit provides insight into how well a topic model captures the underlying structure of a text corpus, both for comparing alternative models and for optimizing internal model parameters. In the context of LDA, model fit is typically evaluated using in-sample and predictive measures. In-sample fit refers to how well the model explains the data on which it was

trained. A common approach is to assess the (log) likelihood of the observed words given the model parameters (e.g., Griffiths et al. 2004, Büschken & Allenby 2016):

$$\log p(w) = \sum_{d=1}^D \sum_{v=1}^V n_{v,d} \cdot \log \left(\sum_{t=1}^T \theta_{d,t} \phi_{t,v} \right) \quad (2.7)$$

Here, $n_{v,d}$ denotes the frequency of term v in document d , $\theta_{d,t}$ represents the proportion of topic t in document d , and $\phi_{t,v}$ denotes the probability of word v under topic t . While the absolute value of the log likelihood is difficult to interpret in isolation, it becomes informative for model comparison or selection – for instance, in identifying the optimal number of topics T or evaluating alternative variants of topic models.

Predictive fit assesses how well the model generalizes to unseen data. It is commonly measured using metrics such as perplexity (e.g., Blei et al. 2003), which are particularly relevant when robustness to unseen documents is essential. Taken together, in-sample and predictive fit offer a comprehensive view of model performance, allowing researchers to assess both the explanatory power and generalizability of LDA-based models.

Topic distribution θ The matrix θ (with dimensions $T \times D$) represents the topic distributions across documents. Specifically, for each document d , LDA generates a multinomial topic probability distribution θ_d , characterizing the document as a mixture of topics. The elements of θ_d sum to one, reflecting the relative proportions of each topic within the document. Figure 2.6 illustrates this idea for the case of $T = 3$ topics. Each point θ_d represents the mixture of topics in a single document. The vertices of the triangle correspond to documents entirely devoted to a single topic, while the points closer to the center represent more heterogeneous mixtures of topics. For example, θ_1 shows a balanced distribution across the three topics, while θ_5 is heavily skewed toward the topic “service experience”. This representation highlights the dimensionality reduction achieved by LDA. Instead of describing documents in the original high-dimensional word space (with vocabulary size V), LDA projects documents into a much smaller T -dimensional topic space, where $T \ll V$. This facilitates interpretable, scalable, and structured analysis of large text corpora, while capturing document-level heterogeneity in topic composition. From a marketing perspective, the set of document-specific topic distributions

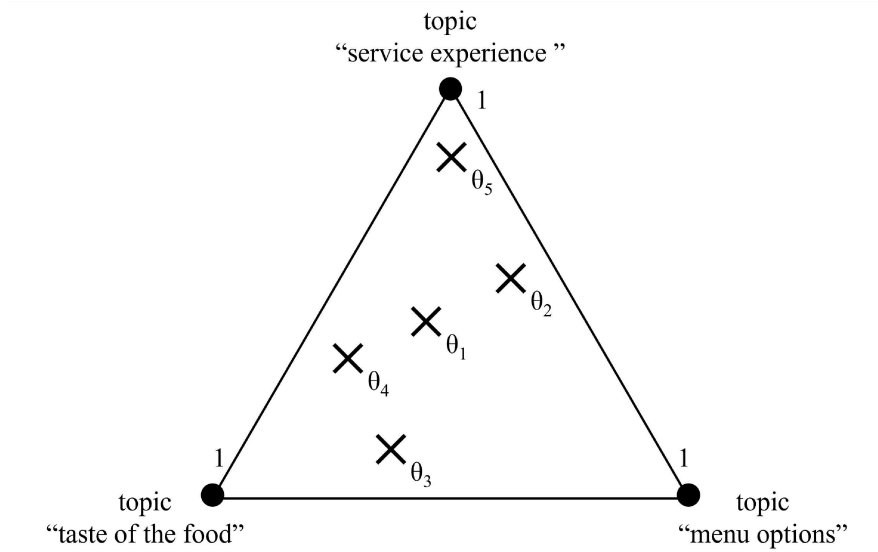


Figure 2.6: Visualization of five documents represented as topic distributions within the topic simplex for the case of $T = 3$ topics. Each point θ_d denotes the topic mixture of a document, where the sum of topic proportions equals one. Adapted from Blei et al. (2003).

$\{\theta_d\}_{d=1}^D$ enables analysts to identify documents, such as customer reviews, that focus heavily on particular aspects (e.g., taste of the food, menu options, or service experience). Furthermore, topic distributions θ can serve as covariates in more advanced analyses. For example, they can be used to predict outcome variables such as consumer ratings (for an integrated model, see Büschken & Allenby 2016).

Word probability distribution ϕ The matrix ϕ (with dimensions $V \times T$) represents the probability distribution over words for each topic inferred by LDA. It is a corpus-level statistic that identifies topics as distributions over terms that tend to co-occur across documents. Accordingly, it is standard practice in topic modeling to interpret and report the most likely words for each topic t . This is accomplished by sorting the entries of the topic-specific probability vector ϕ_t in descending order and mapping the highest-probability values to their corresponding terms. In this way, the top N most likely terms for each topic can be extracted (e.g., Blei et al. 2003, Griffiths et al. 2004, Blei 2012, Büschken & Allenby 2016). Table 2.3 presents three exemplary topics derived from a corpus of consumer reviews of American Restaurants, each represented by its ten most probable terms.

These term lists are typically assigned descriptive labels by researchers, based on their

	“Menu Options”	“Service Experience”	“Social Context”
1	fries	time	husband
2	cheese	\$	wife
3	chicken	about	girlfriend
4	sandwich	waitress	ordered
5	sauce	table	opinion
6	burger	after	family
7	salad	one	brother
8	good	came	date
9	meat	minutes	son
10	barbecue	food	mom

Table 2.3: Three exemplary topics represented by their ten most likely terms, based on a corpus of American Restaurant customer reviews.

interpretation of the shared semantic theme (e.g., Blei et al. 2003, Blei 2012, Büschken & Allenby 2016). In the example, Topic 1 was labeled “Menu Options”, while the others were labeled “Service Experience” and “Social Context”, respectively. This labeling facilitates the qualitative interpretation of LDA outputs, enabling the summarization of large text corpora in a low-dimensional and interpretable format. Beyond qualitative interpretation, various quantitative metrics have been proposed to assess the quality of inferred topics. One widely used measure is topic coherence, which evaluates the extent to which the top words within a topic are semantically related and can be grouped under a common theme. Another important criterion is topic uniqueness, which assesses whether different topics describe distinct aspects of the corpus or exhibit redundancy. The literature offers several operationalizations of coherence (e.g., Mimno et al. 2011, Newman et al. 2010, Bouma 2009, Aletras & Stevenson 2013, O’callaghan et al. 2015) and uniqueness (O’callaghan et al. 2015, Büschken & Allenby 2020), which will be discussed in greater detail in the empirical section of this thesis.

Chapter 3

Model development

In the preceding chapters, we approached the object of investigation, text data, from two distinct perspectives. First, we explored the structure of language, specifically focusing on the syntactic relationships between words within sentences. Our primary goal was to understand “how” language is structured and to investigate how its underlying syntactic organization can be computationally represented, particularly through the Universal Dependencies (UD) framework. Subsequently, we reviewed Latent Dirichlet Allocation (LDA), the most widely used topic model, which serves to summarize “what” text datasets are about. LDA is a stochastic model that conceptualizes text data as multinomial data, treating documents as collections of co-occurring words. Its principal objective is to summarize text data in a parsimonious way by inferring patterns of word co-occurrence at the document level, referred to as topics. In this model, linguistic considerations regarding the structural composition of documents play only a marginal role.

At this point, an inherent discrepancy becomes evident. This stems from the fact that text corpora are not naturally composed of multinomial data but rather consist of documents written in natural language. As discussed above, natural language is rich in structuring elements such as paragraphs, sentences, and grammatical conventions. We adopt the position that a topic model designed to analyze textual data should account for this underlying structure by incorporating additional assumptions grounded in linguistic theory. The notion that language-based documents are merely unordered collections of words – as implied by LDA’s bag-of-words assumption –

overlooks the often intentional and complex structure of natural language. The objective of this chapter is to develop a topic model variant that combines both views on language-based data: it recognizes syntactic dependency structures and utilizes them as complementary information in the process of topic inference.

This chapter is structured as follows: Section 3.1 presents a detailed critique of the bag-of-words assumption inherent in LDA. Section 3.2 reviews prior topic modeling approaches that seek to overcome this assumption by incorporating textual structure. Building on these contributions, we then motivate the development of a syntactically informed topic model. Finally, Section 3.3 provides a formal description of the labeled UD-LDA model.

3.1 Revisiting LDA’s BoW within documents assumption

Our objective is to critically reexamine the premise that LDA is indifferent to the internal structure of text documents. Specifically, LDA does not incorporate any information related to linguistic or textual organization within the boundaries of the document. This limitation stems from the underlying bag-of-words assumption, which treats documents as unordered collections of words drawn from a fixed vocabulary. As a result, word counts at the document level serves as the sufficient statistic for topic inference. The topic indicator z is assigned to each word independently of its local positional context within the document (Figure 2.5). It is therefore inconsequential whether words appear in close proximity, such as within the same paragraph, sentence, phrase, or shared grammatical construction. LDA disregards these structural elements, which implies that topic inference operates exclusively at the global document level. This raises a fundamental question as to whether such an assumption is adequate for capturing the complexity inherent in natural language documents.

In empirical applications, it is rare for textual documents, such as customer reviews, to be presented as unstructured compilations of words. In contrast, such documents typically exhibit a rich array of structural elements. Words are systematically organized into coherent and interpretable phrases and sentences, conforming to (among many others: grammatical) conventions. Moreover, successive sentences are often grouped into distinct textual sections,

such as paragraphs or chapters, each demarcated from others. In many cases, a substantial number of these structural elements are directly observable within the text. When analyzing such documents, the analytical focus shifts from the document level to the local context of individual words. The emphasis is no longer solely on the co-occurrence of words within a document, but rather on the grammatical dependencies between words, their co-occurrence within sentences, or their relative proximities. The central argument advanced here is that adequately capturing the structural complexity of natural language data necessitates greater attention to the specific local contexts in which words appear.

The integration of supplementary structural information – specifically the positioning of words relative to others within a local context – into a probabilistic topic model renders the assumption of conditional independence among words or topics within this context increasingly untenable. Grammatical structures, for example, can influence the likelihood of selecting certain words by assigning low probabilities to grammatically incongruent options in specific syntactic positions. Furthermore, the systematic construction of coherent ideas through sentence formation suggests a potential thematic interdependence among words within the same sentence. Incorporating local structural elements into topic modeling therefore necessitates a more nuanced treatment of word interdependencies within these contexts, along with suitable modeling strategies. The second argument advanced here is that once the analytical focus shifts to the local context, the assumption of conditional independence must be abandoned. Instead, it becomes necessary to formulate explicit assumptions that account for the nature and structure of word interdependencies within local context.

We maintain the position that the fundamental assumption of LDA, which postulates that natural language documents can be treated as a bag-of-words, is excessively simplistic. We advocate for the incorporation of additional information regarding “how” language is structured, thus shifting the analytical focus toward more local contexts within documents. This perspective enables a more precise specification of the assumptions concerning word interdependencies in local contexts. The underlying idea is not new; in fact, a number of prior studies have sought to relax the bag-of-words assumption in LDA in order to more accurately reflect the structural

characteristics of natural language. The following section reviews several of these topic modeling approaches that have been proposed in the academic literature.

3.2 Relaxing the BoW assumption

LDA is frequently used as the foundational framework for advanced topic modeling techniques, where the underlying assumptions are modified or extended by including additional model components. In variants of topic models tailored for text-based datasets, these extensions often involve incorporating information and assumptions regarding the structural composition of the documents under analysis. However, the nature of the supplementary structural features varies across contributions in the literature. These additional elements may be related to word order, sentence organization, grammatical dependencies, or other linguistic features, thereby influencing what is considered the relevant local context. Depending on the model, the local context may encompass the immediately preceding word(s) in a sequence, an entire phrase or sentence, or even the full document.

Furthermore, supplementary assumptions are typically introduced to capture interdependencies between words within the defined local context. For example, the likelihood of observing a specific word may depend on its preceding word, reflecting interdependencies at the word level. Alternatively, interdependencies may arise at the latent topic level, whereby the topic assignments of words are correlated within contextual boundaries. As a result, topic inference becomes intrinsically tied to the observed local context within documents.

To systematically review the various variants of topic models proposed in the literature, the following subsections 3.2.1 to 3.2.3 are organized according to the primary type of **additional structural language features** used for modeling. Although model classifications are not always clear cut, given that some approaches integrate multiple structural features, this categorization facilitates a structured and comprehensive overview. For each model, we also describe how the **local context** is defined, i.e., the contextual scope deemed relevant for topic inference. Finally, we describe the assumed **interdependencies**, indicating how the relationships between words are modeled within the specified context.

3.2.1 Observed sequence of words

A straightforward approach involves conditioning on the observed sequential order of words within documents. This represents an intuitive structural property of texts, given that written language is typically processed in a linear, sequential manner. As such, word order is directly observable in most textual datasets and can be readily exploited as supplementary information for topic modeling.

A widely recognized contribution that incorporates the sequential structure of words within documents is presented by **Wallach (2006)**. This work extends LDA by integrating principles from n-gram models. In n-grams, the probability of observing a specific term is conditioned on the sequence of the preceding n terms, thereby emphasizing the importance of word order. Wallach (2006) incorporates bi-grams into LDA, specifying that the local context for generating a word w_i is determined by the immediately preceding word w_{i-1} . Accordingly, when generating the i -th word in a document, given a topic assignment $t = z_i$, the word is drawn from a distribution over words ϕ_{w_{i-1}, z_i} with $w_{i-1} \in [1, V]$. This results in V distinct distributions, each conditioned on the preceding word w_{i-1} . Consequently, interdependencies at the word level emerge due to the influence of the immediate local context. The contextual scope for sampling topic assignments z_i , however, remains at the (global) document level, with topics independently drawn from θ_d , consistent with the approach in LDA. Wallach (2006) reports improvements in predictive accuracy and identifies topics that are less affected by high-frequency stopwords compared to standard LDA. These findings highlight the value of incorporating structural features of language, such as the sequential order of words, into the topic inference process.

Wang et al. (2007) extend the model introduced by Wallach (2006), based on the observation that successive words (i.e., bi-grams) may form collocations or phrases that convey a distinct meaning beyond the sum of their individual components (e.g., “white house”, Wang et al. 2007). However, such compositional meaning is not consistently observed across all contexts (e.g., “yellow house”, Wang et al. 2007). To account for this variability, the authors augment the bi-gram topic model by introducing a latent “bi-gram status” variable, x_i , which depends on both the preceding word and its associated topic. This bi-gram status variable allows the model to determine whether or not two successive words should be treated as forming a bi-gram. If $x_i = 1$,

the word w_i is drawn from a bi-gram model ϕ_{w_{i-1}, z_i} as in Wallach (2006). If $x_i = 0$, the word is generated from ϕ_{z_i} as in LDA. Their model defines local context as the word immediately preceding in the observed sequence of words. Word interdependencies within this local context are captured through the latent bi-gram status variable.

Additionally, an alternative perspective prevalent in the literature posits that the sequential order of words within documents can offer insights into interdependencies at a functional level. Specifically, this approach suggests that the observed word sequence facilitates the classification of words into distinct categories, typically distinguishing between “functional” and “thematic” words.

The composite model developed by **Griffiths et al. (2004)** is based on the premise that words within documents can be categorized as either syntactic (“functional”) or semantic (“thematic”). The authors argue that syntactic words appear consistently across different documents and serve as structural elements that support the comprehensibility of language. In contrast, semantic words vary depending on the document and are responsible for conveying meaning. The authors extend the LDA framework by integrating a Hidden Markov Model (HMM), resulting in the **HMM-LDA**. The HMM component is designed to capture first-order transitions between latent word classes based on the observed word sequence. Transitions occur from the preceding word’s class c_{i-1} to the focal word’s class c_i , with $c \in [1 : C]$. One of these latent classes $c = 1$ is designated as the semantic class. For words classified as semantic, the model reverts to the LDA mechanism, where a topic assignment z is drawn from θ_d , followed by the word being drawn from the topic-specific distribution $\phi_{t=z}^{(semantic)}$. For all other classes $c \neq 1$, words are drawn directly from class-specific multinomial distributions at the corpus level, denoted $\phi_c^{(syntactic)}$. The HMM component is conditioned on the observed sequential word order, in accordance with the reading direction. Empirical findings indicate that the inferred word classes successfully differentiate between content-bearing and functionally supportive words.

The work of **Darling et al. (2012)** and **Darling & Song (2013)** explores a related area, focusing on the enhancement of unsupervised POS tagging through their proposed model, labeled **Parts-of-Speech LDA (POSLDA)**. The authors adopt a framework that distinguishes between syntactic and semantic words, positing that content words are jointly characterized

by their syntactic class and topic assignment, whereas syntactic words remain independent of topics. Their methodology incorporates both content and functional word classes, modeled using a Hidden Markov Model (HMM). Each functional class is associated with a single word probability distribution $\phi^{(FUN)}$, while each content class is defined by a separate word distribution for every topic, resulting in $K \times S^{(CON)}$ topic-specific content word distributions $\phi^{(CON)}$. By modeling multiple content classes, the authors successfully identify distinct content-related PoS categories, including nouns, verbs, and adjectives. Their findings underscore that linguistic features, particularly the sequential order of words, support the identification of relationships between adjacent words (i.e., the local context) at a functional level. Although the primary objective of POSLDA was to enhance PoS tagging rather than topic inference, the authors report a reduction in perplexity compared to standard LDA, indicating an improvement in model generalizability.

The models discussed thus far demonstrate that the integration of minimal additional structural information – particularly the observed sequence of words – enables the specification of interdependencies among co-occurring words within a local context. This, in turn, yields beneficial effects on topic inference, thereby underscoring the value of structural language features in the context of topic modeling.

3.2.2 Observed sentence boundaries

A commonly utilized linguistic feature is the explicit identification of sentence boundaries, typically indicated by punctuation marks. Accordingly, documents can be conceptualized as sequences of sentences, each of which can be further decomposed into sequences of words. Sentence boundaries have been used as natural cues for topic transitions within documents.

The study conducted by **Gruber et al. (2007)** is based on the assumption that topic inference is influenced by sentence boundaries and the sequential arrangement of words within documents. To account for this, the authors introduce the **Hidden Topic Markov Model (HTMM)**, which models topic transitions between successive words using a Markov Chain mechanism. Specifically, the model incorporates a binary topic transition variable ψ_n for each word w_n . When $\psi_n = 1$, the topic assignment z_n is drawn from θ_d , following the procedure of LDA. Conversely,

when $\psi_n = 0$, the topic is inherited from the preceding word, that is, $z_n = z_{n-1}$. The authors restrict $\psi_n = 0$ to those words that do not initiate a sentence. This implies that topic changes are permissible only at the beginning of a sentence, thus maintaining a consistent topic assignment across all subsequent words within a sentence. The HTMM thus integrates sentence boundaries and word order as additional textual features to improve topic inference. The local context comprises the immediately preceding word in the observed sequence, along with an indicator specifying whether the focal word is the first word of a sentence. All subsequent words within a sentence are deterministically assigned to the same topic as the first word of that sentence. At the beginning of a sentence, a new topic may be introduced, as indicated by the binary topic transition variable $\psi_{begin_sentence} \sim \text{Binom}(\varepsilon)$. Although sentence boundaries are employed as natural points for topic transitions, the authors do not provide a linguistically grounded justification for this modeling assumption. However, empirical results indicate that HTMM improves both perplexity and topic coherence compared to standard LDA.

Sentence boundaries also play a central role in the study by **Büschken & Allenby (2016)**, where the authors argue that consumers usually structure their reviews into sentences. They propose incorporating sentence boundaries into the process of topic inference while simultaneously leveraging customer ratings to inform the distribution of topics across documents. In the **Sentence-Constrained LDA (SC-LDA)**, documents are conceptualized as collections of sentences, with a single topic assigned to all words within a sentence. This constraint ensures that all words in a sentence share the same topic, thereby introducing deterministic interdependencies at the topic level within each sentence. While the order of words within sentences is treated as exchangeable, the model prohibits reordering words across sentence boundaries, a modeling assumption referred to as “bag-of-sentences” (Büschken & Allenby 2016). In a second model variant, termed sticky SC-LDA (with topic carryover), the authors relax the bag-of-sentences assumption by allowing topic carryover between adjacent sentences. Empirical results demonstrate that both sentence-constrained variants outperform standard LDA in terms of predictive fit. Additionally, the topics identified by the SC-models tend to explain customer ratings more effectively than those identified by LDA. Notably, the authors find that the “stickiness” of sentence-based topics is not a dominant phenomenon in the analyzed consumer review data sets.

That is, topic assignments are not frequently carried over from one sentence to the next. In their discussion, the authors further propose that additional linguistic structures present in reviews may provide opportunities to improve their models.

Jiang (2009) extend the distinction between content and functional word states through the introduction of the **nested HMM-LDA** model. This model captures transitions between latent word states within the observed word sequence, while explicitly incorporating sentence boundaries as structural constraints. In line with the assumption that topics are generally consistent within sentences, a single topic is assigned to each sentence. Given the topic of the sentence, latent state assignments are drawn for each word using a Hidden Markov Model (HMM), where the state of each word is conditioned on the state of the preceding word within the same sentence. The model defines a collection of $G = C + F$ latent states for each topic. F denotes the set of functional states characterized by shared word distributions across topics and C refers to the set of thematic states, each defined by a unique topic-specific word distribution. As a result, content words are generated from distributions that reflect both their thematic state and topic assignment, whereas function words are drawn from distributions specific to their functional state, independent of topic. By simultaneously modeling the sequential structure of words and the boundaries between sentences, the approach captures two distinct types of interdependencies within the local context. Sentence boundaries impose interdependencies at the topic level, enforcing a consistent topic assignment across all words within a sentence. Within sentences, the sequential order of words informs state transitions among latent word classes. The model successfully uncovers a correspondence between latent states and grammatical categories: nouns, verbs, and adjectives tend to align with thematic states, whereas pronouns, prepositions, and determiners are typically assigned to functional states.

The models reviewed in this subsection demonstrate that the incorporation of structural information – specifically observed sentence boundaries – facilitates the specification of interdependencies between words within sentences. In particular, interdependencies at the topic level can be clearly defined for words that co-occur within the same sentence, consistent with the intuition that topics tend to remain coherent within individual sentences. Moreover, observable

sentence boundaries serve as natural markers in the text, indicating potential points of topic transition.

3.2.3 (Quasi-) observed grammatical features

Another supplementary textual feature that can inform topic inference is the grammatical structure of texts. However, it is important to note that such structures are not always directly observable. Instead, many grammatical features must be “augmented” during the preprocessing stage prior to topic analysis. Common preprocessing techniques include POS tagging, phrase structure analysis, or dependency parsing. The widespread availability of software tools that support these preprocessing steps allows these grammatical features to be treated as quasi-observed. The existing literature documents a range of efforts aimed at incorporating grammatical structures to enhance topic modeling.

Balikas et al. (2016) introduce **copLDA**, a model designed to capture topic interdependencies within textual segments using copulas. The approach is based on the assumption that text segments, such as sentences or noun phrases, exhibit thematic coherence. While sentence boundaries are directly observable features of text, the identification of noun phrases necessitates an additional preprocessing step. In the two proposed model variants, topic correlations are introduced within local textual boundaries, specifically (1) sentences and (2) noun phrases. Within these segments, copLDA introduces topic-level dependencies via copula functions. Consequently, the assignment of topics to words within text segments is conditioned on both the marginal probabilities θ_d and a copula distribution. The authors employ a Frank copula for the purpose of modeling interdependencies between topic assignments within text segments, introducing an additional hyperparameter λ . A value of $\lambda = 0$ corresponds to complete independence, while $\lambda = \infty$ indicates perfect dependence of topic assignments within text segments. When $\lambda = 0$ and the segment is defined as the entire document, the model reduces to standard LDA. The authors empirically evaluate various λ values across datasets and identify optimal values greater than zero for both variants of the model, which supports the assumption of correlated topics within segments. Notably, the optimal λ is higher for noun phrases ($\lambda = 5$) than for sentences ($\lambda = 2$), indicating stronger thematic coherence within noun phrases. Perplexity-based evalua-

tions demonstrate improved model fit over LDA, particularly for the noun phrase variant, and yield topic sequences that align more closely with human judgment. These findings underscore the utility of incorporating (quasi-)observed grammatical structures into topic inference.

Büschken & Allenby (2020) propose the **Autocorrelated Topic Model (AT-LDA) with covariates**. This model enables first-order topic carryover from one word to the next within the observed word sequence. In addition, it incorporates grammatical textual features – such as sentence conjunctions and punctuation – to inform the probability of topic carryover.

Topic carryover is modeled via a latent binary indicator $\zeta_n \sim \text{Binomial}(\psi_n | z_{n-1})$. A value of $\zeta_n = 1$ implies that the topic assignment is carried over from the preceding word ($z_n = z_{n-1}$). Conversely, if $\zeta_n = 0$, the topic assignment z_n is drawn from the document-level topic distribution θ_d , in line with classical LDA. The topic carryover probability ψ_n is specified through a binary logit model. It includes observed grammatical features as covariates and allows the intercept to depend on the topic assignment of the previous word z_{n-1} . Thus, the probability of topic carryover is a function of both local grammatical cues and topic assignments. This requires knowledge of word order to identify features such as conjunctions and punctuation.

The local context is characterized by the elements immediately preceding in the observed text sequence. This encompasses not only the immediately preceding word but also considers the preceding structural elements of text, such as conjunctions or punctuation. The interdependence of words within that local context is conceptualized through probabilistic modeling of topic carryover.

In their empirical analysis of customer review datasets, Büschken & Allenby (2020) provide evidence of local serial topic dependencies, referred to as “topic runs”. The probability of topic carryover is significantly influenced by structural linguistic elements, particularly sentence conjunctions and punctuation. The authors report improvements in topic inference. These improvements become more pronounced in more complex datasets, such as those with longer documents or larger vocabularies. In such settings, both in-sample and predictive fit outperform standard LDA. Topic quality, measured through coherence and uniqueness, is also enhanced under greater textual complexity. In addition, topics inferred by AT-LDA with covariates exhibit higher explanatory power for customer ratings than those derived from standard LDA.

The AT-LDA model is particularly relevant for this thesis for two main reasons. First, the model and its empirical results highlight that grammatical language features shape the latent topic structure in texts. This supports the view that structural aspects of language (*how* texts are written) interact with topical content (*what* is represented). Second, the model is technically attractive due to its flexibility and extensibility. Any (quasi-)observed textual feature at the word level can be integrated as a covariate in the binary logit specification of topic carryover. This opens the door for incorporating more complex grammatical features beyond sentence conjunctions and punctuation.

Büschken & Allenby (2025) propose the **Topic Change Point Model (TCPM)**, which represents text as a sequence of topic runs. These are sequential word segments originating from the same topic and potentially interspersed with ubiquitous, non-topical terms. The model enforces a topic change from one run to the next, with the boundary between runs defined as topic change points. These change points are unobserved and inferred from the data. A variant of TCPM incorporates observed Parts-of-Speech (POS) covariates to inform change point inference, linking syntactic categories to the likelihood of changes in the topic.

A distinctive feature of TCPM is the use of observed POS tags, i.e., syntactic classifications of words, alongside the word sequence as structural language features. The authors show empirically that certain POS categories are systematically associated with topic shifts: for example, sentence boundaries and coordinating conjunctions often indicate a topic change, whereas determiners and personal pronouns tend to indicate topic continuation. This demonstrates how grammatical features can inform the detection of topic boundaries, capturing the interplay between syntactic and thematic structure in texts.

In TCPM, local context is defined at the phrase level, with phrases corresponding to the sequence of words between two change points. These contextual boundaries are not predefined (as with sentence boundaries) but are inferred. Within each phrase, topic words share a single topic assignment, while non-topic words function as “*syntactic glue*” that ensures readability but does not convey topical content. Modeling these non-topic words removes the need for manual stopword pruning, as the model identifies such words automatically based on their context-independent frequency. Empirical results show that TCPM facilitates topic inference

using characteristic topic phrases rather than isolated word lists and achieves superior fit to observed data compared to other topic models.

Boyd-Graber & Blei (2008) incorporate (quasi-) observed syntactic dependency structures into the **Syntactic Topic Model (STM)**. The observed data consists of a collection of documents, each represented as a set of dependency-parsed sentence trees. These dependency structures are treated as observed textual features and must be derived by preprocessing of the raw text. Each sentence is represented by **unlabeled** dependency trees, characterized by hierarchical parent-child relationships between words. The term **unlabeled** indicates that the grammatical type of each dependency relation is not specified. The authors propose a nonparametric topic model in which the number of topics is inferred from the data, rather than fixed in advance, as in LDA. The model conditions on dependency tree structures to generate topic assignments z for words. These assignments are informed by parent-child relationships within the tree structure. Specifically, the topic assignment for a child word is drawn from a distribution defined as the product of two vectors, θ_d and $\pi_{z_{parent}}$:

$$z_{child} \propto \theta_d \cdot \pi_{z_{parent}} \quad (3.1)$$

where θ_d denotes the topic proportions of document d , and π is a corpus-level topic transition matrix. This matrix encodes the transition probabilities from parent topics to child topics. Importantly, topic transition probabilities are assumed to be invariant across documents and types of dependency relations. The vector $\pi_{z_{parent}}$ specifies the transition vector, given the topic assignment of the parent word. As a result, the model introduces topic-level interdependencies: the topic assignment of a child word is conditioned on the topic assignment of its syntactic parent. This creates a chain of interdependent topic assignments along the branches of the dependency tree.

From a linguistic perspective, the methodology employed by the authors is particularly relevant to this study, as it incorporates syntactic dependency structures into the topic model. This approach extends the representation of textual data beyond simple word sequences and sentence boundaries by establishing parent-child relationships between words, relationships that

do not necessarily correspond to their linear order in the text. In the empirical application of their model, the authors used the Penn Treebank (Marcus et al. 1993), a corpus composed of manually curated constituent parse trees. Subsequently, these constituent trees were converted into dependency parse trees (Johansson & Nugues 2007). Although the conversion process produces labeled dependency trees, Boyd-Graber & Blei (2008) chose to disregard dependency labels, focusing exclusively on the tree structure of unlabeled dependencies. We contend that the use of dependency labels merits closer examination. Our hypothesis is that interdependencies at the topic level are influenced not only by the existence of parent-child relationships, but also by the grammatical distinctions encoded in these relationships.

The literature reviewed in this section has thus far treated grammatical structures as quasi-observed phenomena. In addition, several studies attempt to jointly infer syntactic structures and latent topics. Although not the central focus of this thesis, such approaches represent efforts to exploit the interaction between grammatical characteristics and topical content.

For example, Wallach (2008) introduces a Bayesian model for dependency parsing. This generative framework requires joint inference of both latent dependency tree structures and the associated latent states. These latent states can be interpreted as “finer-grained POS tags” or “syntactic topics”. For instance, the model is capable of generating multiple states that exhibit a tendency to produce nouns. While one of these states generates nouns related to job titles, another state is more likely to produce nouns denoting place names. In her model, the emphasis is on inferring the surface structure of dependency trees, with no emphasis on labels for different types of relationships. In contrast, our approach assumes that labeled dependency trees are available through appropriate preprocessing. Moreover, our focus lies on investigating how syntactic dependency labels contribute to topic inference.

The reviewed studies indicate that grammatical features of text can support topic inference. Structural elements such as noun phrases, sentence conjunctions, punctuation, and dependency trees have been effectively utilized to model word interdependencies at the topic level.

Authors	Model	Structural Features Incorporated	Scope of Local Context	Modeled Interdependencies
Wallach (2006)	Bigram Topic Model	Sequential order of words	Preceding word(s) in sequence	Serial word-choice interdependencies
Wang et al. (2007)	Topical N-gram Model	Sequential order of words	Preceding word(s) in sequence	Serial word-choice interdependencies (bigram status)
Griffiths et al. (2004)	HMM-LDA	Sequential order of words	Preceding word in sequence	Serial word-class (syntactic/semantic) interdependencies
Darling et al. (2012), Darling & Song (2013)	POSLDA	Sequential order of words	Preceding word in sequence	Serial word-class (syntactic/semantic) interdependencies
Gruber et al. (2007)	HTMM	Sequential order of words and sentence boundaries	Preceding word in sequence	Sentence-level topic uniformity and serial topic interdependencies between sentences
Büschen & Allenby (2016)	(sticky) SC-LDA	Sentence Boundaries (and sequential order of sentences)	Words within the same sentence	Sentence-level topic uniformity (and serial topic interdependencies between sentences)
Jiang (2009)	nested HMM-LDA	Sequential order of words and sentence boundaries	Preceding word in sequence within sentence	Sentence-level topic uniformity; serial word-class (syntactic/semantic) interdependencies
Balikas et al. (2016)	copLDA	Sentence boundaries OR noun phrase (NP) boundaries	Words within the defined boundaries (sentence or NP)	Sentence-level or NP-level topic interdependencies
Büschen & Allenby (2020)	AT-LDA	Sequential order of words, sentence conjunctions and punctuation	Preceding textual elements in sequence	Topic-level serial interdependencies
Büschen & Allenby (2025)	TCPM	Sequential order of words and parts-of-speech (POS) tags	Sequence of words in between unobserved topic change points	Phrase-level topic uniformity, interspersed by non-topic words
Boyd-Graber & Blei (2008)	Syntactic Topic Model	Unlabeled dependency trees	Syntactic parent-child dependency pairs	Topic-level interdependencies along dependency tree branches

Table 3.1: Overview of topic modeling approaches incorporating (observed) structural features of text data.

3.2.4 Summary

The outlined approaches share the common characteristic of incorporating (in)directly observed information on “how” language-based text data is structured. In addition to document boundaries (Blei et al. 2003), natural language offers a range of structural features that serve as informative cues for topic modeling. For instance, many models condition on the observed word sequence (Griffiths et al. 2004, Wallach 2006, Wang et al. 2007, Darling et al. 2012) or additionally account for the presence of sentence conjunctions and punctuation within that sequence (Büschken & Allenby 2020). Sentence boundaries are another frequently used structural feature (Gruber et al. 2007, Jiang 2009, Büschken & Allenby 2016). Moreover, some approaches leverage indirectly observed syntactic structures that require additional preprocessing steps, such as dependency trees (Boyd-Graber & Blei 2008) or noun phrases (Balikas et al. 2016).

The observed structural features (or their combinations) define a **local context** for topic modeling. This local context may comprise the n preceding elements in a sequence, the set of words within a sentence, or syntactically connected words. Conditioned on such local context, models can specify interdependencies between linguistic elements that go beyond simple word co-occurrence. These interdependencies manifest at different levels: the word level (e.g., Wallach 2006, Wang et al. 2007), the topic level (e.g., Boyd-Graber & Blei 2008, Balikas et al. 2016, Büschken & Allenby 2016, 2020) or the word-function level (e.g., Griffiths et al. 2004, Darling et al. 2012, Jiang 2009). The nature of these interdependencies ranges from deterministic to probabilistic or conditionally independent. A summary of the relevant literature is provided in Table 3.1.

We interpret the reviewed approaches as a synthesis of two perspectives on text data: the question of “how” text is structured and the question of “what” content is represented. Observed information on “how” language is structured supports the analysis of “what” latent topical content is conveyed. The literature demonstrates that both perspectives can be combined in a complementary manner. Specifically, (in)directly observed structural features of language allow for the definition of a local context. Given this context, interdependencies between linguistic elements can be modeled. This, in turn, facilitates a more accurate derivation of the latent topic structure within text data.

We see further opportunities for research that draws more explicitly on syntactic features grounded in linguistic theory. So far, grammatical features in topic modeling have largely been limited to sentence boundaries, conjunctions, noun phrases, or dependency tree structures. However, none of these approaches explicitly considers the **type of syntactic relationships between words**. The following section addresses this gap by introducing a novel topic model variant that incorporates labeled syntactic dependency structures. Our intended application in marketing research further underscores the relevance of this approach. In many marketing contexts, the availability of data is limited, making efficient use of available information especially important. From our perspective, this includes making use of syntactic features inherent in natural language.

3.3 Labeled Universal Dependencies LDA

This section integrates the perspectives developed in Chapter 2.1 and Chapter 2.2 into a unified framework: the **labeled Universal Dependencies LDA (labeled UD-LDA)**, a novel topic model variant based on labeled syntactic dependency structures. The central premise of this study is that understanding the syntactic structure of language (*how* language is constructed) can support the inference of topics, or *what* is expressed in a given text. Our approach builds on prior work in the topic modeling literature, as reviewed in Chapter 3.2, which has demonstrated the value of integrating structural features of text into topic inference. In the following section, we discuss how three key contributions have informed the development of the labeled UD-LDA: the foundational LDA model by Blei et al. (2003), the AT-LDA model by Büschken & Allenby (2020), and the STM model by Boyd-Graber & Blei (2008).

Foundations and limitations of existing approaches As discussed earlier, the foundational topic model LDA (Blei et al. 2003), introduced in Section 2.2, incorporates minimal structural information specific to text data. It conditions solely on document boundaries, which define the scope for topic inference. This is known as the bag-of-words assumption within documents.

However, this assumption contrasts with the general observation that texts are both formally and thematically structured. Typically, texts present a coherent sequence of ideas, with words

grouped according to thematic relevance. For instance, consider the restaurant review sentence: “*The senior menu is a very good bargain*”. Here, the two words “*senior menu*” collectively express a specific offering, while “*very good bargain*” describes its great value. The words are thematically interdependent, collectively expressing a meaningful proposition.

Büschken & Allenby (2020) address this issue by conditioning on the observed word sequence. Based on the local context within this sequence, their model allows for autocorrelated topics, where topics may be inherited from one word to the next. Topic propagation offers an elegant and flexible way to capture local interdependencies between words. Figure 3.1 (b) illustrates potential topic pathways through such propagation (shown in blue).

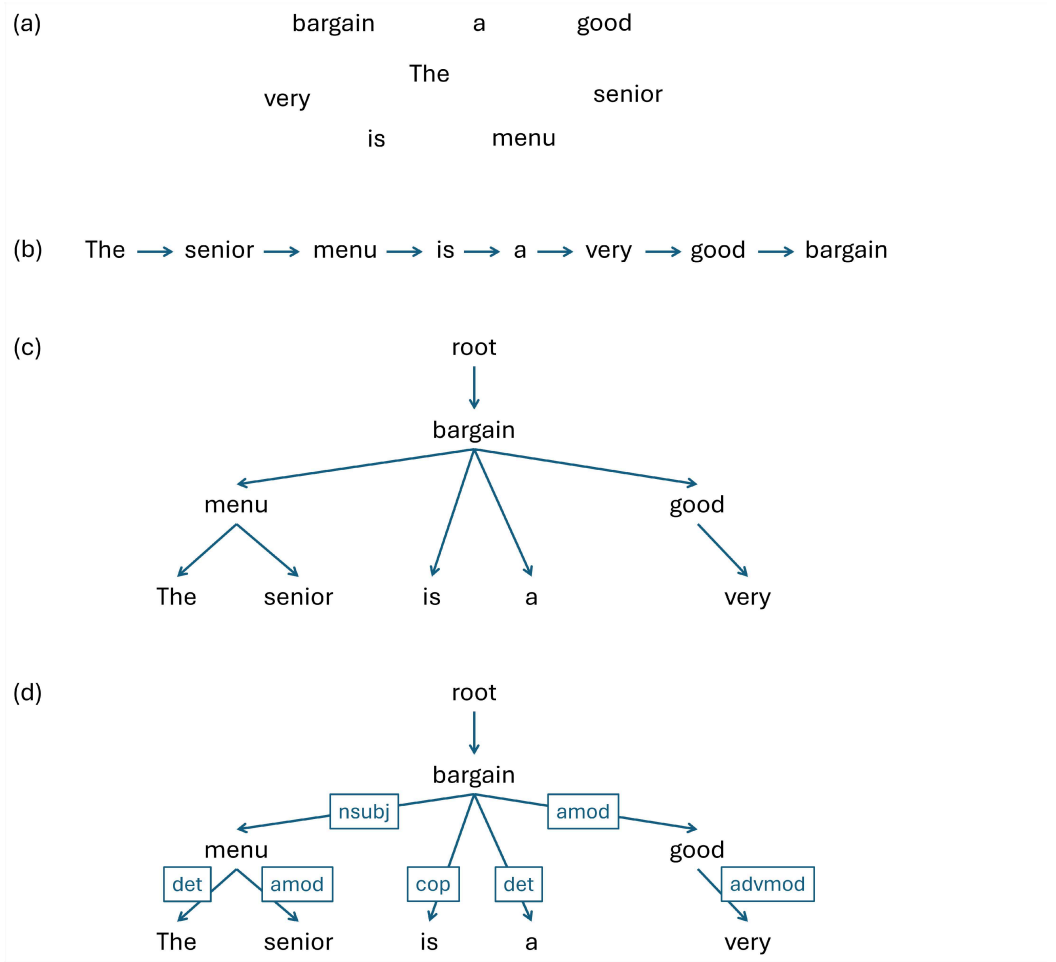
We argue, however, that topic propagation does not always proceed linearly through the word sequence. In particular, retaining stopwords complicates a plausible interpretation of topic inheritance.¹ It is questionable, for example, why the word “the” should transfer its topic to “senior”. Intuitively, the direction should be reversed. Similarly, it seems unlikely that “senior” transfers its topic to “menu” – rather, “menu” is the thematic anchor, with “senior” specifying it. Hence, the reading direction aligns only partially with the intuitive flow of thematic dependencies.

Figure 3.1 (c) shows the same sentence represented as a parsed UD tree, with dependency labels omitted for clarity. A closer look at the blue arrows reveals that they align with intuitive notions of topic inheritance. For example, the word “menu” may pass its topic to the words “the” and “senior”. Similarly, “bargain” may propagate its topic to the word “good”, which further qualifies the nature of the deal. These patterns suggest that UD tree structures can reflect and formalize intuitive thematic relationships between words. They thus provide a linguistically grounded foundation for modeling topic flow at the sentence level.

The integration of syntactic tree structures into topic modeling is not new. Boyd-Graber & Blei (2008) introduced the Syntactic Topic Model (STM), which relies on observed syntactic parse trees. Here, a word’s local context is defined by its syntactic head (upstream) and its dependents (downstream). Topic interdependencies are represented as topic transitions along tree branches rather than through topic carryover. The position of words within parse trees determines the local context, which is comprised of the word that is the head of the focal word (“upstream”)

¹As discussed in the introduction, the amount of data is inherently limited in targeted marketing analyses. For this reason, we advocate against aggressive data pruning during preprocessing.

Figure 3.1: Illustration of sentence structure representations and the potential paths for topic propagation.



Notes. (a) A sentence in bag-of-words representation. (b) A sentence represented in the observed sequence of words. (c) A sentence in syntactic dependency representation showing hierarchical relationships among words. (d) A sentence in Universal Dependency representation with labeled dependencies. The blue arcs indicate possible paths for topic propagation, demonstrating how different structural representations of a sentence can impact topic propagation.

and the words that are in turn dependents of the focal word (“downstream”). However, STM assumes that the topic transition probabilities π_k are invariant with respect to the type of head-dependent relation. Conditional on the topic assignment of the head word $z_h = t$, the probabilities of topic transitions from head to dependent are independent of their type of syntactic relation.

Figure 3.1 (d) illustrates the same sentence with UD labels included. These labels help identify which syntactic relationships are likely to support topic propagation: For example, modifier relations such as “amod” and “advmod” tend to preserve topic continuity, suggesting a high probability of topic propagation from head to dependent. Notably, the verb “is” does not serve as the root of the sentence. Instead, it functions as a copula attached to “bargain”. The *cop* relation marks “is” as a function word with limited autosemantic content, subordinated to its more contentful head “bargain”. This reflects a core principle of UD: content-bearing words are preferred as syntactic heads over grammatical function words (De Marneffe et al. 2021). Accordingly, “is” may inherit its topic from “bargain”. These examples show that labeled UD trees can formalize thematic relationships that are intuitively perceived in natural language. In light of the findings by Büschken & Allenby (2020), who demonstrate that grammatical cues like conjunctions and punctuation affect topic propagation, we argue that dependency labels offer additional grammatical information relevant for topic modeling.

To the best of our knowledge, no existing topic modeling approach integrates syntactic tree structures together with head-dependent labels. We address this gap by proposing labeled UD-LDA. The central idea of labeled UD-LDA is to use (observed) syntactic structures from labeled dependency trees for inferring topic interdependencies in text.² Labeled UD-LDA models autocorrelated topics by allowing topic propagation along the branches of UD trees. The probability of topic propagation is defined as a function of both the topic of the head word and the type of syntactic relationship between head and dependent.

Labeled UD-LDA A graphical representation of the labeled UD-LDA is illustrated in Figure

3.2. The left-hand side of the figure shows the UD structure for a parsed sentence. Each

²In practical applications, topic modeling typically involves substantial preprocessing of raw text. This includes sentence segmentation, tokenization, and often stopword removal or stemming. UD parsing is a natural extension of this preprocessing pipeline and can be implemented with minimal additional effort using available software. As such, we treat parsed dependency structures, including their grammatical labels, as (quasi-) observed pseudo-data.

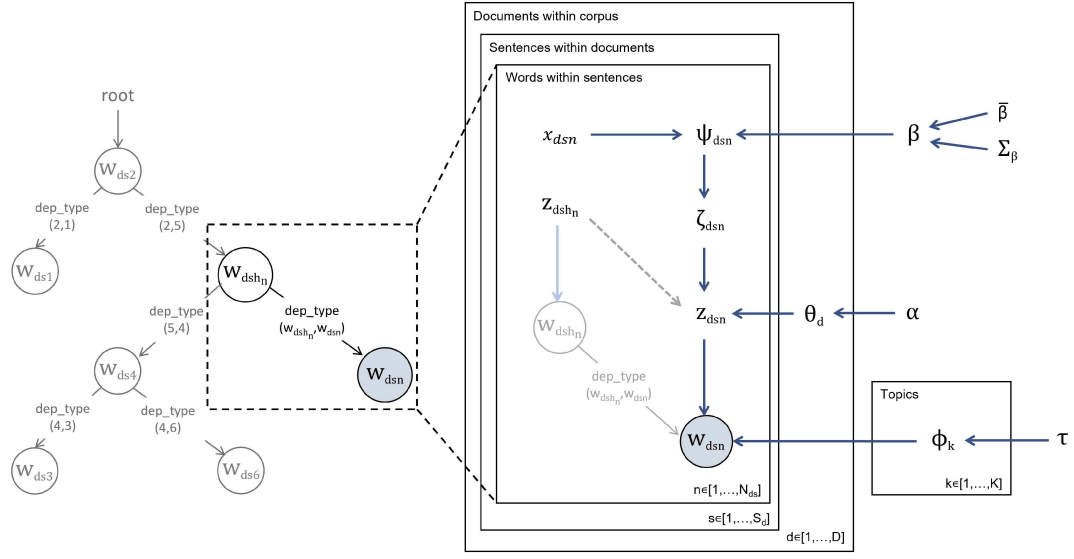


Figure 3.2: Labeled Universal Dependencies LDA (labeled UD-LDA).

node represents a word in the sentence, and labeled, directed edges indicate head-dependent relationships. The sentence contains a single root word that serves as the initial point of connection. All other words are linked to the root through a unique path within the syntactic tree. This structure provides insight into the grammatical relationships between words in a sentence. Directed edges identify which word pairs (w_{dsh_n}, w_{dsn}) are in direct head-dependent relationship. The dependency label assigned to each edge, $dep_type(w_{dsh_n}, w_{dsn})$, specifies the grammatical nature of the relationship. By treating this syntactic structure as observed, sentence boundaries define a local context for modeling topic interdependencies within documents. This is reflected in the “*sentences within documents*” plate on the right-hand side of Figure 3.2. Within each sentence, the labeled UD-LDA further iterates over “*words within sentences*”. The focal word is denoted as w_{dsn} (highlighted in gray), and its unique syntactic head is w_{dsh_n} .

The topic assignment z_{dsn} of the focal word is generated using the topic propagation mechanism introduced by Büschken & Allenby (2020). A latent binary variable ζ_{dsn} determines whether the topic of the focal word is inherited from its syntactic head. This variable is the key component that allows for autocorrelated topics within the labeled UD-LDA framework.

$$\begin{aligned} \zeta_{dsn} = 1 : z_{dsn} &= z_{dsh_n} \\ \zeta_{dsn} = 0 : z_{dsn} &\sim \text{Multinomial}(\theta_d) \end{aligned} \quad (3.2)$$

In the first case, $\zeta_{dsn} = 1$, the topic of the focal word is inherited from its syntactic head. This results in a deterministic topic propagation: $z_{dsn} = z_{dsh_n}$. In this case, the link between z_{dsn} and the document-level topic distribution θ_d is deactivated. Instead, the topic assignment of the syntactic head, z_{dsh_n} , is directly passed to the focal word.

In the second case, $\zeta_{dsn} = 0$, no topic propagation occurs from the syntactic head. Here, the relationship between z_{dsh_n} and z_{dsn} is disabled. The topic assignment for the focal word, z_{dsn} , is then drawn from a multinomial distribution parameterized by the document-specific topic proportions θ_d .

In the lower part of the labeled UD-LDA model, the generation of the focal word w_{dsn} is conditioned on its topic assignment $z_{dsn} = t$, regardless of whether the topic was assigned via propagation or sampling from θ_d . The word is drawn from a topic-specific multinomial distribution: $w_{dsn} \sim \text{Multinomial}(\phi_{k=t})$.

What remains to be discussed is the origin of the binary topic propagation indicator ζ_{dsn} . Following the mechanism proposed by Büschken & Allenby (2020), we assume that ζ_{dsn} is drawn from Bernoulli distribution with success probability ψ_{dsn} .

$$\begin{aligned} \zeta_{dsn} &\sim \text{Bernoulli}(\psi_{dsn}|x_{dsn}) \\ \psi_{dsn}|x_{dsn} &= \frac{\exp[x_{dsn}^\top \beta]}{1 + \exp[x_{dsn}^\top \beta]} \end{aligned} \tag{3.3}$$

As specified in Equation 3.3, the probability of topic propagation, ψ_{dsn} , is modeled as a function of word-level covariates x_{dsn} and corpus-level regression coefficients β . The labeled UD-LDA differs from the AT-LDA (Büschken & Allenby 2020) in the definition of the covariate vector x_{dsn} . Here, the covariates are constrained by Universal Dependencies in two ways. First, UD specifies the syntactic head of each word. This allows conditioning on the topic assignment of the head word. Second, the UD label provides a grammatical specification of the relationship between the head and the dependent. Thus, x_{dsn} is constructed as the concatenation of two dummy-coded components: (1) the topic assignment of the head word, z_{dsh_n} , and (2) the dependency relation type, $dep_type(w_{dsh_n}, w_{dsn})$:

$$x_{dsn} = \begin{bmatrix} z_{dsh_n} \\ dep_type(w_{dsh_n}, w_{dsn}) \end{bmatrix} \quad (3.4)$$

The resulting vector x_{dsn} has length $K + J$, where K denotes the number of topics and J the number of UD relation types (as defined in De Marneffe et al. 2021). Correspondingly, β is a vector of corpus-level regression coefficients of length $K + J$, consisting of two parts: The first K elements are topic-specific intercepts β_z , and the remaining J elements are dependency-type-specific coefficients β_{dep_type} . Positive values of β increase the probability of topic propagation, while negative values reduce it. The topic-specific intercepts allow for differences in propagation propensity across topics, consistent with the findings in Büschken & Allenby (2020). The coefficients associated with dependency types enable the model to capture systematic variation in topic propagation behavior based on grammatical structure. The data generating process can thus be described as follows:

1. Draw independently:
 - (a) Draw ϕ_k from $\text{Dirichlet}(\tau) \quad \forall k$ i.i.d.
 - (b) Draw θ_d from $\text{Dirichlet}(\alpha) \quad \forall d$ i.i.d.
 - (c) Draw β from $\text{MVN}(\bar{\beta}, \Sigma_\beta)$
2. For each sentence $s \in [1 : S_d]$ in each document $d \in [1 : D]$:
 - (a) For the root word w_{root} :
 - i. Draw z_{root} from $\text{Multinomial}(\theta_d)$
 - ii. Draw w_{root} from $\text{Multinomial}(\phi_{k=z_{root}})$
 - (b) For non-root words w_n with $n \in [1 : N_s], n \neq root$:
 - i. Compute $p(\psi_n | x_n, \beta, z_{h_n})$ using (3.3) and draw ζ_n from $\text{Bernoulli}(\psi_n)$
 - ii. If $\zeta_n = 0$, draw z_n from $\text{Multinomial}(\theta_d)$; if $\zeta_n = 1$, set $z_n = z_{h_n}$
 - iii. Draw w_n from $\text{Multinomial}(\phi_{k=z_n})$

By the definition of the UD framework, each sentence contains exactly one root word (De Marneffe et al. 2021). This root word is the only word in the sentence that has no head but serves as

its own head. As a result, the topic assignment of the root word, z_{root} , cannot arise from topic propagation and is always drawn from θ_d . This also implies that the labeled UD-LDA does not model topic propagation across sentence boundaries. Furthermore, it is important to emphasize that the syntactic structure of sentences is not inferred by the model. Instead, the UD structure is treated as observed and incorporated into the design matrix of the topic propagation regression. Thus, the syntactic UD structure serves as a fixed skeleton that is subsequently populated with correlated topic assignments and words.

The joint distribution of knowns and unknowns of the labeled UD-LDA model, given document d , and assuming independent prior distributions, factorizes as follows.

$$\begin{aligned}
 p(\{w\}_d, \{z\}_d, \{\zeta\}_d, \phi, \theta_d, \beta, x_d, \alpha, \tau, \bar{\beta}, \Sigma_\beta) \propto \\
 \prod_{s=1}^{S_d} \left[p(w_{s,root} | \phi, z_{s,root}) \times p(z_{s,root} | \theta_d) \times \right. \\
 \left. \prod_{\substack{n=1 \\ n \neq root}}^{N_s} \left[p(w_{s,n} | \phi, z_{s,n}, z_{s,h_n}, \zeta_{s,n}) \times p(z_{s,n} | z_{s,h_n}, \theta_d, \zeta_{s,n}) \times p(\zeta_{s,n} | x_{s,n}, \beta, z_{s,h_n}) \right] \right] \quad (3.5) \\
 \times p(\phi | \tau) \times p(\theta_d | \alpha) \times p(\beta | \bar{\beta}, \Sigma_\beta) \\
 \times p(\tau) \times p(\alpha) \times p(\bar{\beta}) \times p(\Sigma_\beta)
 \end{aligned}$$

Equation 3.5 identifies the parameters that require sampling to approximate their posterior distributions and estimate the model. Given fixed priors $\tau, \alpha, \bar{\beta}, \Sigma_\beta$, sampling is required for $\phi, \theta_d, \beta, z, \zeta$. Büschken & Allenby (2020) developed an Markov Chain Monte Carlo (MCMC) sampling scheme for estimating the AT-LDA topic model with covariates. This scheme can be adapted with minor modifications for use in the labeled UD-LDA. A version of the algorithm tailored to the labeled UD-LDA is provided in Appendix B of this thesis.

(Formal) relation to existing topic modeling variants The labeled UD-LDA can be formally related to the three topic modeling variants discussed earlier in this section.

LDA (Blei et al. 2003) can be regarded as a special case of labeled UD-LDA. As previously outlined, the case $\zeta_n = 0$ disables the relationship between z_{h_n} and z_n , resulting in z_n being drawn from a multinomial distribution with document-specific topic probabilities θ_d . This implies that z_n is conditionally independent of all other topic assignments in the document. This mechanism is identical to the topic assignment procedure in classical LDA. Hence, LDA corresponds to the special case of labeled UD-LDA in which $\zeta_n = 0$ for all $n \in N_d$. Under this restriction, all topic assignments in a document are sampled from $Multinomial(\phi_{k=z_n})$, and word generation $w_n \sim Multinomial(\phi_{z_n})$ proceeds as in LDA.

The labeled UD-LDA is also technically related to the **AT-LDA** model developed by Büschken & Allenby (2020). Both models allow autocorrelated topic assignments via probabilistic topic propagation. However, the labeled UD-LDA introduces additional structural assumptions. First, topic propagation is restricted to occur only within sentence boundaries; there is no propagation across sentences. Second, topic propagation is permitted only from a word’s syntactic head w_{dsh_n} to its dependent w_{dsn} . In AT-LDA, topic propagation is defined sequentially, with inheritance from w_{dn-1} to w_{dn} . This can be viewed as a special case of syntactic dependency where $h(n) = n - 1 \quad \forall n \in 2 : N_d$. The transition from serial dependency to syntactic (Universal) dependency requires monitoring the positions of words and their corresponding head within text documents. In AT-LDA, the document’s “tree” structure forms a single branch aligned with the reading order (see Figure 3.1 **b**). By extending this with labels indicating sentence conjunctions and punctuation, AT-LDA can be embedded as a special case of the labeled UD-LDA.

Only limited formal parallels exist between labeled UD-LDA and the **STM** (Boyd-Graber & Blei 2008). A key difference is that STM is a nonparametric model, allowing the number of topics to be inferred from the data, whereas labeled UD-LDA assumes a fixed number of topics. Both models treat the input data as a collection of dependency parse trees. However, STM does not incorporate dependency labels. Its topic transition probabilities π_k are not conditioned on syntactic relation types. In contrast, labeled UD-LDA uses dependency labels as covariates in the topic propagation mechanism. The models also differ in their mechanisms for generating

topic assignments. In STM, the topic assignment is derived from the normalized product of the document-specific topic distribution and the topic transition probabilities ($z_n \propto \theta_d \pi_{z_{h(n)}}$). In labeled UD-LDA, topic assignments results either from propagation ($z_n = z_{h(n)}$) or from a draw from θ_d , depending on the outcome of the binary variable ζ_n . Given the substantial differences in model architecture and assumptions about topic interdependencies, no formal mapping exists that would render one model a special case of the other.

Chapter 4

Empirical analysis

4.1 Data

The empirical implementation of the labeled UD-LDA is guided by a marketing-oriented perspective. This necessitates its evaluation within the context of a targeted analysis. The objective of this targeted analysis is to identify latent topics in customer reviews with the (general) purpose of elucidating aspects of consumer behavior such as the origin of product evaluations or recommendations. Unlike large-scale, undifferentiated text corpora, marketing analyses are typically based on targeted data sets that reflect a specific intersection of key dimensions such as brand, product type, price level, distribution channel, and consumer segment. Accordingly, the more precise objective of this empirical application is to uncover latent topics that explain consumer perceptions and behaviors in distinct market contexts. These insights can support marketers in designing and refining offerings that resonate with their target groups.

Three targeted data sets of publicly available customer reviews are analyzed in this chapter. Each data set corresponds to a specific marketing context and exemplifies a realistic scope for applied marketing research. The first dataset consists of 1,683 reviews of American style restaurants (collected from we8there.com). This data set reflects the intersection of geographic location, cuisine type, and price level. The second dataset includes 3,210 reviews of luxury hotels in Manhattan (collected from expedia.com). This collection reflects the intersection of brand, premium price positioning, and geographic location. The third dataset comprises 7,772

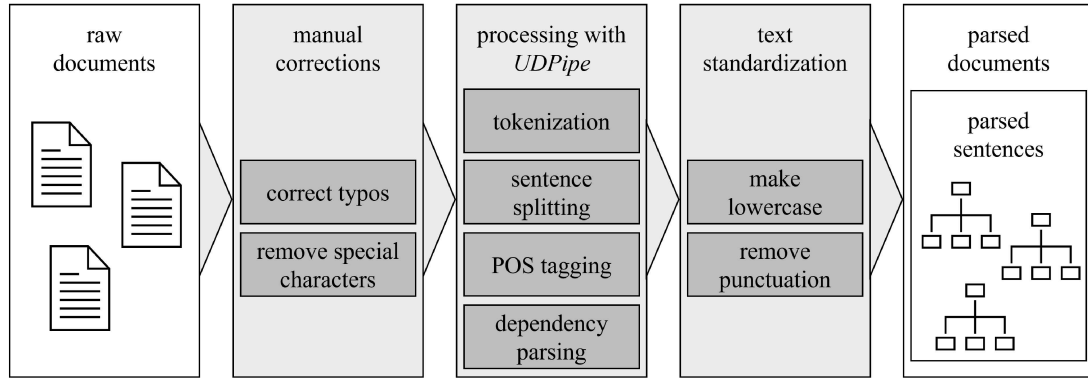


Figure 4.1: Text data pre-processing for the labeled UD-LDA.

customer reviews of camping tents (collected from amazon.com). It reflects the intersection of product category, price level, and online distribution channel. The relatively small size of all three corpora, each ranging from several thousand to fewer than 10,000 reviews, is consistent with the scope of targeted marketing analyses. As discussed in the introduction, this aligns with the constraints of real-world marketing settings, where data sets are often limited in size yet highly specific in focus.

A three-step preprocessing pipeline was applied to the text data prior to the empirical analysis (see Figure 4.1). These steps included (1) manual corrections, (2) dependency parsing using *UDPipe*, and (3) text standardization. The first step involved manual corrections of the raw text. This included the removal of spurious characters (e.g., HTML tags, encoding artifacts) and the correction of frequent spelling errors (e.g., “Manhatan”, “Manhattan”). In contrast to typical preprocessing routines, stopwords, infrequent terms, and punctuation were retained, and capitalization was left unchanged. These choices were made to preserve the original sentence structure to the greatest extent possible, thereby facilitating accurate dependency parsing. For example, preserving the capitalized form of proper nouns or the pronoun “I” supports correct syntactic analysis in English. At the end of this step, each review was represented as a single character string.

In the second step, the character strings were passed to the *UDPipe* toolkit (Straka et al. 2016, Straka & Straková 2017), which performs sentence segmentation, tokenization, POS tagging, lemmatization, and dependency parsing. We used a pre-trained model based on the English Web Treebank (EWT) corpus (Silveira et al. 2014) via the *UDPipe* R wrapper (Jan Wijffels 2022).

The EWT corpus is viewed as a gold-standard UD corpus comprising over 16,000 sentences from genres such as weblogs, emails, newsgroups, reviews, and Yahoo! answers. The inclusion of the latter two categories, both user-generated content, motivated our use of a model trained on this corpus. At the end of this processing step each document is divided into individual sentences that are tokenized into words and punctuation marks. Each token is annotated with Universal Dependency information. Of particular interest for our model are the ID of each token's head and the type of dependency relation with its head. This information enables the representation of each sentence as a typed dependency tree.

The third and final preprocessing step involved text standardization. At this stage, lowercasing was applied, and punctuation marks were removed. Punctuation is excluded from topic modeling, as it carries no semantic content related to latent topics. This removal does not interfere with the UD tree structure, since punctuation tokens are always terminal nodes and never serve as syntactic heads. A common practice in topic modeling is to eliminate infrequent terms to reduce vocabulary size (e.g., Blei et al. 2003, Griffiths et al. 2004, Jiang 2009, Büschken & Allenby 2020). This study deliberately avoids removing rare words or stopwords. Doing so could disrupt the dependency tree structure by eliminating the head of other dependents, thereby invalidating the syntactic relationships central to our analysis. Since this research focuses on modeling topic propagation along syntactic paths, preserving these structures is essential.

In conclusion, although the preprocessing pipeline may seem elaborate, it introduces only minimal alterations to the original text. Its primary objective is to preserve the linguistic structure while systematically incorporating UD information.

Automated UD parsing of raw text is not error-free. Therefore, it is essential to consider the potential limitations of this approach. The complexity of automatic UD parsing is well recognized in the NLP community. The Conference on Computational Natural Language Learning (CoNLL) highlighted this challenge by making UD parsing its shared task in both 2017 and 2018 (Zeman et al. 2017, 2018). Standard evaluation metrics include the Labeled Attachment Score (LAS), which measures the percentage of tokens correctly assigned both a syntactic head and a dependency label, and the Unlabeled Attachment Score (UAS), which measures only head assignment accuracy (Zeman et al. 2017, 2018). For English gold-standard

corpora, *UDPipe* 2.5 achieves UAS scores between 80.2% and 83.9% and LAS scores between 75.9% and 80.4%.¹ While these figures may appear modest, they are competitive by current standards in UD parsing. Alternative tools such as Stanza, implemented in Python, report slightly higher accuracy (UAS = 86.2, LAS = 83.6 on the English-EWT corpus) (Qi et al. 2020). However, integrating Stanza would complicate the workflow, as the present project is implemented in R. Manual correction of UD structures, although theoretically ideal, is infeasible for two reasons. First, inter-annotator agreement on the English-EWT corpus is reported at only 96%, indicating that even human annotators exhibit some degree of inconsistency.² Second, manual corrections are highly time and cost intensive, which is impractical for large-scale analyses. A core premise of this study is that syntactic dependency information can be used for downstream tasks with minimal effort. The manual inspection and correction of parsed trees would contradict this principle. Furthermore, our interest does not lie in dependency structures themselves, but in their contribution to downstream tasks such as topic modeling. Accordingly, we assess the utility of automatically parsed UD structures for topic inference, as they are realistically available to applied researchers using publicly accessible tools such as *UDPipe*.

An overview of the descriptive statistics of the datasets after preprocessing is presented in Table 4.1. Although all three datasets fall within the same genre of customer reviews, they differ in both size and complexity. The American Restaurant dataset, despite containing the fewest reviews, shows the most extensive vocabulary and the longest average review lengths of approximately 191 words (or 14 sentences). Luxury Hotel reviews are the shortest, averaging approximately 53 words (or 4.3 sentences) per review. Reviews of considerable length, ranging between 575 and 1,699 words, are found across all datasets. While the datasets differ at the document level, these differences diminish considerably when attention is shifted to the sentence level. The average sentence length is nearly uniform across all datasets, ranging from 12.4 to 14.4 words per sentence. These findings suggest that consumers tend to formulate sentences of comparable complexity when describing products or services. Longer reviews arise not from longer sentence constructions but from a higher number of sentences.

¹https://ufal.mff.cuni.cz/udpipe/1/models#universal_dependencies_25_models

²https://github.com/UniversalDependencies/UD_English-EWT/blob/master/README.md

Statistic	Luxury Hotels	Camping Tents	American Restaurants
Number of reviews	3,210	7,772	1,683
Corpus size	169,389	682,494	321,451
Number of unique terms	6,019	10,521	11,966
Number of words per review			
Mean	52.8	87.8	190.9
Standard deviation	46.9	111.4	225.8
Median	40	52	108
Maximum	575	1,697	1,699
Number of sentences per review			
Mean	4.3	6.1	14.0
Standard deviation	3.1	6.5	15.5
Median	4	4	9
Number of words per sentence			
Mean	12.4	14.4	13.6
Standard deviation	8.4	9.2	8.0
Median	11	13	12

Table 4.1: Descriptive statistics of the three customer review datasets. Summary statistics for the Luxury Hotels, Camping Tents, and American Restaurants datasets, including corpus size, vocabulary size, and distributional measures for words and sentences per review.

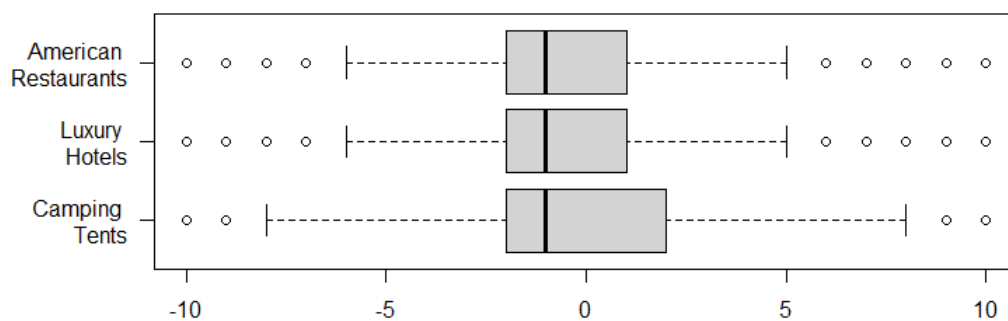


Figure 4.2: Distributions of head-dependent distance across datasets. Head-Dependent distance is measured as the difference in word positions within a sentence. Positive values indicate that the dependent follows its head in the observed word sequence (dependency relation oriented in reading direction), whereas negative values indicate that the dependent precedes its head (dependency relation oriented against reading direction). A value of +1 indicates that the dependent immediately follows its head. A distance of zero occurs for sentence roots, which are their own heads. The vertical bar marks the median value of each distribution.

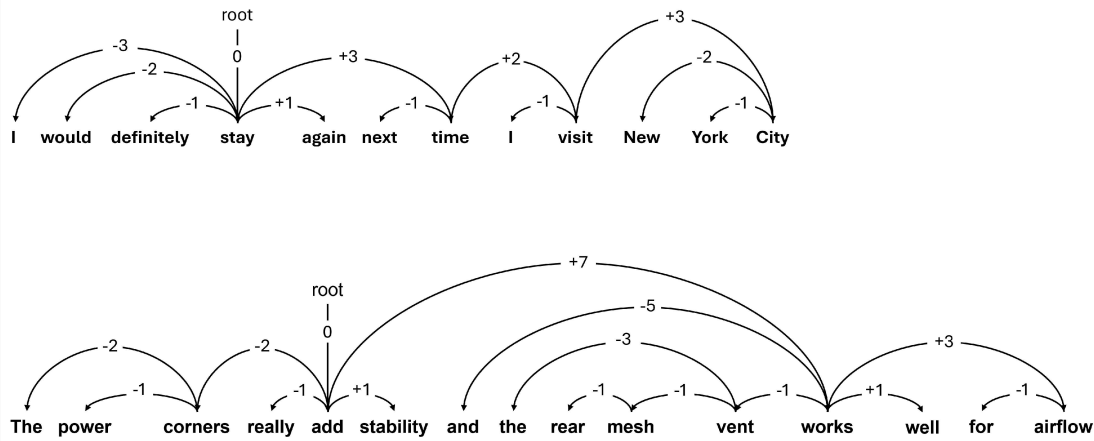


Figure 4.3: Head-dependent distances in parsed sentences. Two example sentences are shown, the first from the Luxury Hotels dataset and the second from the Camping Tents dataset. Arrows indicate head-dependent relations, with numeric labels showing the head-dependent distance. Positive (negative) values indicate that the dependent follows (precedes) its head in the observed word sequence, while a value of zero marks the sentence root, which is its own head.

In addition to sentence length, UD parsing enables the analysis of dependency distances, defined as the difference in word positions between a head and its dependent within a sentence. A value of zero is reserved for sentence roots, which are their own heads; thus, the proportion of zero distances in a dataset equals the number of sentences. Positive values indicate that the dependent follows its head in the observed word order (i.e., the dependency relation is oriented in reading direction), while negative values indicate that the dependent precedes its head (i.e., the relation is oriented against reading direction).

Figure 4.2 summarizes these distributions across the three datasets. A consistent pattern emerges: the majority of dependencies are concentrated around short distances, with the lower quartile at -2 and the upper quartile at +1 for the American Restaurants and Luxury Hotels datasets, and at +2 for Camping Tents. This clustering around zero confirms that syntactic dependencies are overwhelmingly local, linking words that are close together in the linear sequence. At the same time, the distributions show a substantial spread, with many dependencies extending over several intervening words. Moreover, the median lies at -1 for all datasets, indicating that more than half of the dependencies run against the reading direction, with dependents preceding their heads in linear sequence. This shows that reverse-direction dependencies are not marginal exceptions but a central characteristic of the syntactic structures observed in these datasets.

Figure 4.3 provides qualitative illustrations: in both example sentences, some dependencies connect immediate neighbors (e.g., *stay* $\rightarrow$ *again*), while others span larger gaps (e.g., *add* $\rightarrow$ *works*, distance +7). In addition, several relations run against the reading direction (e.g., *would* $\rightarrow$ *stay*, distance -3), indicating that the dependent often precedes its head.

Taken together, these observations highlight three key characteristics of syntactic structure. First, dependencies are predominantly local, showing that the relevant context for most words is typically located within a short range in the same sentence. Second, non-adjacent dependencies are not rare anomalies but a systematic property of sentence structure, indicating that the relevant context of a word often extends beyond its immediate neighbors. Third, the prevalence of reverse-direction dependencies shows that many dependents precede their heads, meaning that relevant conditioning information frequently lies to the right of a token in the linear sequence.

These findings carry important implications for topic modeling. Models that rely solely on sequential proximity cannot account for the non-adjacent and bidirectional nature of syntactic dependencies. As discussed in Chapter 3.2, extensions of LDA often assume first-order interdependencies in linear sequence: some model first-order word choice probabilities (Wallach 2006, Wang et al. 2007), others model transitions between word classes (Griffiths et al. 2004, Jiang 2009, Darling et al. 2012, Darling & Song 2013), or topic propagation across neighboring words (Büschken & Allenby 2020). These approaches demonstrate that incorporating local context can improve topic inference. However, their left-to-right perspective, while a convenient modeling assumption and one that aligns with human reading practices, overlooks fundamental aspects of how language organizes context.

More recent distributional approaches explicitly acknowledge that relevant context is both bidirectional and often extends beyond adjacent words. Word2Vec, for example, exploits the two-sided local context (CBOW predicts a word from both the left and right neighbors; Mikolov et al. 2013). BERT generalizes this idea with deep bidirectional attention, predicting masked tokens using information from both sides (Devlin et al. 2019). Transformer architectures further extend the context window, enabling the modeling of long-range dependencies (far) beyond sentence boundaries (Vaswani et al. 2017). For autoregressive language generation, however, transformer decoders restrict each position to attending only to earlier tokens, ensuring that predictions

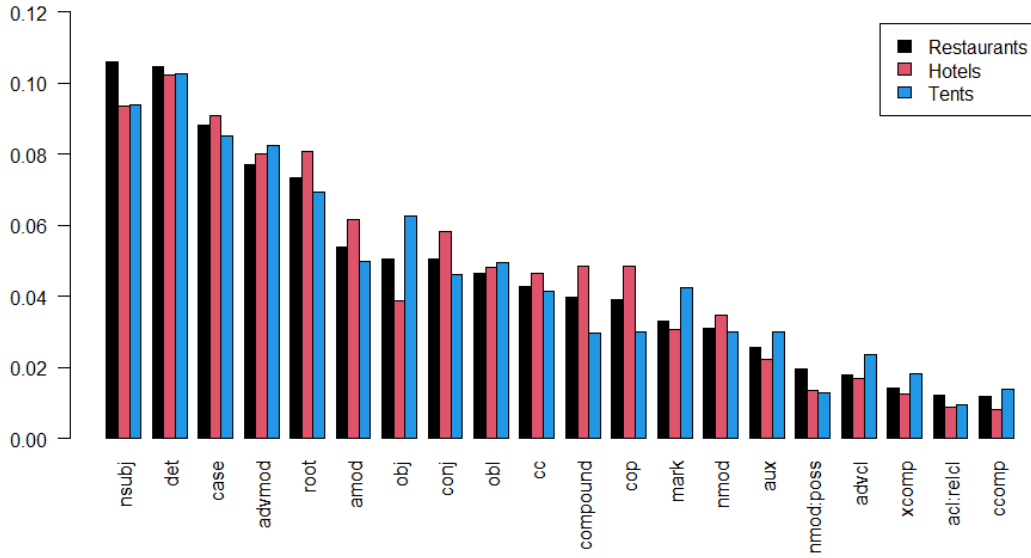


Figure 4.4: Share of most common UD types across datasets.

condition solely on left context (Vaswani et al. 2017). This means that dependencies where the head follows its dependent – such as the relation *corners* → *power* in Figure 4.3 – cannot be represented directly. Transformer models can (successfully) approximate such relationships, but the architecture itself enforces a one-sided, left-to-right conditioning. These models identify relevant context in a largely data-driven manner, relying on large training corpora and highly parameterized architectures, but without explicit guidance from linguistic theory.

Dependency structure offers a theory-informed alternative. It encodes both adjacent and longer-range relations, including cases where dependents precede their heads, thereby providing a principled representation of context that sequential models systematically overlook. UD in particular specifies which word pairs are connected and by what relation type. In light of the empirical distributions above – showing predominantly local links, frequent non-adjacent relations, and many reverse-direction dependencies – UD captures precisely the kinds of contextual connections that first-order sequential assumptions miss. This makes dependency information a promising and interpretable signal for topic inference, especially in targeted, data-scarce marketing settings.

While distance distributions highlight the positional properties of syntactic relations, dependency labels specify the functional role that connects a head to its dependent. These labels may

differ in their informativeness for topic inference: some reflect close semantic refinement, while others mark looser structural connections. Figure 4.4 displays the frequency distributions of dependency types for the three data sets analyzed. Once again, the distributions appear highly similar across datasets. The dependency relations of nominal subjects (*nsubj*), determiners (*det*), and case markers (*case*) are consistently observed to be the most frequent among all data sets. Modifying relations such as adverbial modifiers (*advmod*) and adjectival modifiers (*amod*) are also prevalent. As discussed earlier, these relations may contribute valuable information for topic modeling, as modifiers help specify the meaning of their head words. Nonetheless, some variation in dependency type frequency is observable. For example, the tent dataset exhibits higher frequencies of direct objects (*obj*), subordinating clause markers (*mark*), and adverbial clause modifiers (*advcl*). These patterns suggest a more complex sentence structure in the tent reviews, which aligns with the longer average sentence lengths in this dataset. In summary, the internal syntactic structures of sentences show a high degree of consistency across all three datasets. This structural regularity justifies further investigation into syntactic features as a basis for topic modeling. If syntactic dependencies align with thematic dependencies in text, then observed Universal Dependencies may serve as valuable prior information for inferring latent topic structures.

4.2 Topic inference

This section addresses the first research question **RQ1** of this thesis: *Does a topic model that conditions topic propagation on UD-based syntactic dependencies yield an enhanced representation of the latent topic structure in text datasets?* To this end, we evaluate the labeled UD-LDA’s performance in an empirical application using the datasets described above. First, we introduce the benchmark models (Section 4.2.1). We then compare models using multiple evaluation criteria: model fit (Section 4.2.2), face validity of inferred topics (Section 4.2.3), topic uniqueness (Section 4.2.4), and topic coherence (Section 4.2.5).

4.2.1 Benchmark models

A natural benchmark is **classical LDA**, which does not incorporate any structural features of the underlying text documents beyond word counts. This makes LDA a variant based on the bag-of-words assumption (Figure 3.1 (a)). As previously discussed, LDA can be implemented as a special case of labeled UD-LDA by setting $\zeta_n = 0, \forall n \in N_d$.

Since word sequence and sentence boundaries are commonly used as structural features in topic modeling, appropriate benchmark models should account for this type of information. The **Sentence-Constrained LDA (SC-LDA)** developed by Büschken & Allenby (2016) serves as a benchmark that conditions on observed sentence boundaries while ignoring the sequence of words within sentences. In SC-LDA, all words within a sentence are assigned the same topic, i.e., topic assignments are homogeneous at the sentence level. This model can be viewed as a special case of labeled UD-LDA in which topic propagation is deterministically set to $\zeta = 1$ for all (non-root) words within a sentence.

The next benchmark model additionally incorporates word order within sentences as a structural covariate. We implement a variant of the **Autocorrelated Topic Model (AT-LDA)** developed by Büschken & Allenby (2020). In our implementation, we allow for probabilistic topic propagation between consecutive words within sentences, but do not use sentence conjunctions and punctuation as covariates. Hence, topic propagation is solely conditioned on the observed word sequence within sentences. The AT-LDA is implemented as a special case of labeled UD-LDA, where each sentence is represented as a sequential “dependency tree”. Specifically, the first word $w_{n=1}$ of each sentence is treated as the root, and each subsequent word w_n (for $n \neq 1$) is the dependent of its immediate predecessor w_{n-1} . This results in an unlabeled tree with a single branch, following the observed reading direction (Figure 3.1 (b)). The word-level covariate vector, x_n includes only the dummy-coded topic assignment of the preceding word, z_{n-1} .

A fourth benchmark model aims to incorporate (observed) syntactic head-dependent relationships rather than linear word order. We refer to this variant as **unlabeled UD-LDA**. This model is a special case of labeled UD-LDA that conditions on syntactic dependency trees without using dependency type labels (Figure 3.1 (c)). The only difference from labeled UD-LDA lies in the

specification of the covariate vector x_n , which in this case includes only the dummy-coded topic assignment of the syntactic head word z_{h_n} , omitting the information about dependency type.

All benchmark models, like most topic models, require the number of topics to be specified as an input parameter. In the present study, the number of topics is fixed at $K = 10$ across all experiments. This decision is based on prior research involving the same datasets, which suggests that ten topics sufficiently capture the thematic structure in each of the three corpora. It is also consistent with the targeted nature of the datasets under investigation. Although model performance could be further evaluated by varying the number of topics, fixing $K = 10$ ensures comparability and consistency between model variants.

It is important to emphasize that the objective of the following empirical analyses is not to delineate specific boundary conditions under which UD-based topic modeling outperforms alternative methods. Rather, the analysis aims to explore the potential contribution of syntactic structures to topic modeling. Specifically, we examine whether, and in what way, labeled head-dependent relations influence topic propagation and topic discovery within text data.

4.2.2 Model fit

In this section, we compare the (in-sample) model fit of the proposed labeled UD-LDA to that of the benchmark models described above. While the model fit of a single model on a given dataset is not inherently meaningful in isolation, it serves as a basis for evaluating the relative performance of different models applied to the same data.

As a baseline, we report the model fit for the Unigram, a model-free benchmark that estimates the marginal probabilities of words as $p(w_i) = \frac{c(w_i)}{W}$, where $c(w_i)$ denotes the number of times a specific term w_i appears in the corpus, and W is the total number of words in the corpus. The Unigram corresponds to LDA with a single topic. Thus, a comparison between the Unigram and classical LDA highlights the added value of modeling documents as mixtures of latent topics.

We then turn to models that incorporate observed structural features of text. Specifically, we consider the SC-LDA to assess the contribution of observed sentence boundaries. A comparison with the AT-LDA allows us to examine the added benefit of incorporating sequential word order within sentences to model topic propagation.

We next consider models that incorporate (observed) syntactic tree structures obtained from UD parsing. Comparing AT-LDA and the unlabeled UD-LDA highlights the value of conditioning on syntactic head-dependent relations rather than on sequential word order. Finally, we assess the contribution of grammatical dependency labels by comparing the unlabeled UD-LDA with the labeled UD-LDA. This allows us to assess the contribution of the grammatical type of head-dependent relations in modeling topic propagation.

	Luxury Hotels	Camping Tents	American Restaurants
Unigram	-6.02	-6.05	-6.44
LDA	-5.25	-5.36	-6.11
SC-LDA	-5.64	-5.71	-6.16
AT-LDA	-5.49	-5.50	-6.22
unlabeled UD-LDA	-5.30	-5.08	-5.95
labeled UD-LDA	-5.10	-4.87	-5.86

Table 4.2: Model fit as the log marginal likelihood of the data (per word).

Table 4.2 presents the model fit results for all models variants across the three datasets. It shows the log marginal likelihood of the data, normalized by the total number of words, thereby allowing for comparability across datasets of varying lengths. Across all datasets, the labeled UD-LDA achieves the highest model fit, indicating that incorporating syntactic covariates – specifically labeled Universal Dependency relations – systematically improves the model’s ability to explain the observed data. This finding underscores the benefit of conditioning on grammatical structure when modeling topic propagation, as it captures not only what is said but also how it is linguistically constructed. Notably, the labeled UD-LDA consistently outperforms the classical LDA in all three datasets, despite imposing structural constraints on topic assignment. This is particularly relevant given that classical LDA, under the bag-of-words assumption, remains more flexible by design.

A more nuanced pattern emerges among the remaining benchmark models. Among those incorporating topic propagation, the unlabeled UD-LDA performs better than the AT-LDA across all datasets. This suggests that topic inheritance is more effectively modeled through syntactic head-dependent relations than through word adjacency alone. This finding aligns with our intuitive expectation, as illustrated by the example in Figure 3.1. However, additional structural information does not universally lead to improved fit. In fact, the classical LDA outperforms

both the SC-LDA and the AT-LDA in all three datasets. Even the unlabeled UD-LDA, which utilizes syntactic tree structure without dependency labels, shows slightly lower fit than classical LDA in the Luxury Hotels dataset. These findings suggest that structural constraints like homogeneous sentence-level topics (SC-LDA) or sequential topic carryover (AT-LDA) can limit model flexibility and do not result in improved in-sample fit. This result contrasts with previous findings obtained on the same datasets; however, the present study applied a different preprocessing pipeline, which may partly explain the divergence.

It is important to note that preprocessing choices in this study differ from those in Büschken & Allenby (2016, 2020). Specifically, our approach retains all stopwords and infrequent terms, consistent with our objective to preserve sentence structure for syntactic parsing. Moreover, our implementation of the AT-LDA omits grammatical covariates such as sentence conjunctions and punctuation, which were used in Büschken & Allenby (2020) to inform topic propagation. Under different preprocessing conditions, such as stopword removal or shorter average sentence lengths, sequential topic carryover or homogeneous sentence-level topic assignments may still offer a plausible approximation of topic interdependencies in text.

4.2.3 Face validity of inferred topics

This section evaluates the labeled UD-LDA by assessing the face validity of its inferred topics. To do so, we compare its results with those of classical LDA and validate both models against intuitive expectations derived from a manual inspection of the datasets. Specifically, this involves reading a random sample of reviews from each dataset to gain an understanding of the dataset and the topics discussed by customers. This qualitative evaluation acts as a preliminary sanity check to determine whether the labeled UD-LDA effectively captures topics relevant to the corpus.

Topics are commonly represented by the most probable words from the topic-specific word distributions ϕ_k for each topic $k \in K$. To estimate these distributions, we computed the posterior mean over the final 100 MCMC draws of ϕ . This avoids dependence on a single chain state while remaining computationally efficient.

Across all three datasets, we observe a relatively close alignment between the topics inferred

by the labeled UD-LDA and those from classical LDA. To illustrate this, we examine the Luxury Hotels dataset in more detail. Table 4.3 shows the top 20 words per topic from this dataset for labeled UD-LDA, while Table 4.4 reports the same for classical LDA. We find that both models generally identify interpretable topics, with some overlap in content. For example, both models yield topics that can be interpreted as relating to **location**, **staff**, and **positive overall experience**. These topics not only emerge across both models but also exhibit similar word compositions, suggesting that both approaches capture core thematic patterns in the data.

For instance, consider the **location** topic: in labeled UD-LDA, this corresponds to **Topic 10**; in classical LDA, it corresponds to **Topic 4**. Among the top-ranked words are “close”, “walking”, “distance”, “subway”, “Times”, and “Square” for both models. This topic reflects that guests frequently describe the hotels as centrally located, near key attractions such as Times Square or Central Park. Importantly, this topic aligns with one of the key dimensions for which the dataset was originally collected. As discussed above, the Luxury Hotel reviews represent a targeted corpus defined by the intersection of brand, premium price positioning, and location. The emergence of a clearly interpretable “location” topic therefore provides additional evidence that the inferred topics capture relevant and context-specific themes within the dataset.

Similarly, both models identify a **staff** topic – **Topic 1 in labeled UD-LDA** and **Topic 8 in LDA**. Words like “staff”, “friendly”, “helpful”, and “clean” suggest that reviewers often highlight the hospitality and performance of hotel employees.

Both models also generate a broader topic that we interpret as representing a **positive overall experience**. This corresponds to **Topic 4 in labeled UD-LDA** and **Topic 6 in LDA**. The top words for this topic span multiple aspects of the hotel experience, including terms like “location”, “view”, “experience”, “restaurant”, and “food”. While these terms collectively suggest a positive stay, they are not as closely related as in more specific topics, indicating that this topic may reflect an overall sentiment rather than a narrowly defined theme. Example reviews with high proportions of this topic include: *"Great hotel, excellent location, helpful and friendly staff"*. *"Had a great time, hotel in a great location, service and food excellent."* These reviews are typically short and list various attributes that contributed to a favorable experience.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
and the was very staff room clean hotel were helpful friendly comfortable rooms nice great location service are good is	for our with my in the a to of and at on room us day night stay me check out	the was and is a breakfast not it that but only no there free very room good which so service	hotel the a this was great location is it view perfect place time good stay excellent an experience wonderful nice	to and I we you have not they nt but had if that when did could was it get be	a the room was little bit elevators no were an elevator bed king lot problem floor small rooms be which	we I had were stayed I was have enjoyed loved expected wanted thought they found happy there needed arrived impressed	the in of at Times Square on hotel and to with from floor is room time all Hilton street for	would stay again I definitely recommend will here there we back highly go be staying not return come certainly I	to the and New York in for of a City from close subway is walking all Square Times distance it

Table 4.3: Top 20 words in **labeled UD-LDA** on Luxury Hotels dataset

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
the we and was our were a to of had at in room for it with us hotel stay on	a room and the was with bed in nice very but comfortable on beds bathroom not for small of	New York to in City stay this will hotel for and I my time again at definitely would location everything	to and the hotel a from close of Times walking Square subway walk location for distance central is great station	the was I it a that have hotel but not nt of room to would this in at did only	great location view service hotel excellent time square restaurant wonderful amazing fantastic staff everything experience perfect food an and the	the is you of in and a are to hotel Square Times it if this on right for be all	and was very the staff hotel clean helpful stay were location great nice again room would rooms definitely	for breakfast is and a good but no not free in rooms there with \$ or nice small only	the to and was I in they my room for that check not had when on at desk a were

Table 4.4: Top 20 words in **classical LDA** on Luxury Hotels dataset

While minor differences exist in the highest-probability words across models, the overall findings support the face validity of the inferred topics. The topics appear to reflect genuine content within the corpus rather than artifacts introduced by modeling assumptions.

Despite the substantial overlap between the models, the labeled UD-LDA identified two topics in the Luxury Hotel reviews that were not detected by classical LDA. These topics are particularly worth highlighting, as they may hold managerial relevance. First, **labeled UD-LDA's Topic 6** includes terms such as “elevator(s)” and “problem” (see Table 4.3). This topic, uniquely uncovered by the UD-LDA, reveals issues with the elevators. A manual review of the corresponding customer reviews confirms that elevators were frequently cited as a pain point. Guests complained about their small size, slow performance, or complete malfunction. This is an operational issue that hotel managers could address to improve overall guest satisfaction. The second distinct topic, **labeled UD-LDA's Topic 9**, appears to capture guest intentions to recommend the hotel or return for future stays – indicative of customer loyalty and word-of-mouth in marketing terms. However, the top words in this topic do not explicitly reveal the underlying reasons for such positive evaluations. To identify these drivers, managers could use the per-document topic distribution θ_d to filter and analyze reviews with a high proportion of this topic. This approach may yield actionable insights into what aspects of the experience generate loyalty or positive word-of-mouth, as guests often articulate their reasoning in narrative form.

It is important to emphasize that the discovery of managerially relevant topics ultimately depends on the dataset at hand. The model itself does not guarantee the emergence of actionable topics. Nevertheless, our qualitative evaluation shows that labeled UD-LDA is capable of uncovering such themes when they are present in the data. While the topics inferred by labeled UD-LDA largely align with those identified by classical LDA, nuanced differences in topic composition and clarity emerged – particularly in topics less distinctly captured by classical LDA.

One further observation from the top word lists is that, similar to classical LDA, the topics inferred by labeled UD-LDA are dominated by terms with high marginal probabilities, regardless of topic assignment (e.g., “a”, “be”, “do”, “it”, etc.). As discussed above, there are different strategies to handle such high-frequency words. Some modeling approaches regarded them as function words (e.g., Griffiths et al. 2004, Jiang 2009) while others even treat them as stopwords

to be removed during preprocessing (e.g., Blei et al. 2003). In the labeled UD-LDA, however, these words are not treated differently than other words – they are permitted to inherit the topic of their syntactic head. Regardless of whether these topic assignments result from topic propagation or are sampled from θ_d , they inform the topic-word distributions ϕ and consequently appear among the top-ranked words across topics. This can hinder interpretability. A post-estimation solution is to exclude high-frequency words from the top-word lists – under the assumption that they contribute little to the semantic content of a topic. This filtering step can reduce noise and enhance the interpretability of inferred topics.

What remains open is a systematic evaluation of the topic quality produced by labeled UD-LDA relative to the benchmark models. A direct qualitative comparison across ten topics, five models, and three datasets would require $10 \times 5 \times 3 = 150$ individual assessments, rendering a manual evaluation infeasible. Therefore, the next subsections employ established quantitative measures of topic quality, such as Topic Uniqueness and Topic Coherence, to facilitate a structured comparison of model performance.

4.2.4 Topic uniqueness

The analysis of Topic Uniqueness serves as a quality criterion for evaluating topic models, specifically their ability to generate well-differentiated topics within a given dataset. Ideally, each topic should identify a distinctive set of terms that captures a unique semantic context. The capacity to detect distinct topics is essential for the model’s interpretability and its utility in downstream tasks. However, topic uniqueness is not inherently guaranteed by the model architecture. Rather, it emerges from the interaction between the model and the underlying dataset. It is therefore also an empirical question to what extent a given model, applied to a particular dataset, produces well-separated topics.

Various methods have been proposed in the literature to operationalize topic uniqueness. For example, O’callaghan et al. (2015) assess uniqueness using the mean pairwise Jaccard similarity among the top N most probable words in each topic. This approach yields a single aggregate score per model, quantifying the average overlap across all topic pairs. While informative, such

an average can obscure important differences, as it compresses a potentially wide distribution of pairwise topic overlaps into a single summary statistic.

An alternative approach, adopted by Büschken & Allenby (2020), evaluates uniqueness by explicitly calculating pairwise topic overlap. For each pair of topics i and j in a model with K topics, overlap between the N most probable words is calculated as:

$$\text{overlap}(i, j) = \frac{|\{w_i\} \cap \{w_j\}|}{N} \quad (4.1)$$

where $\{w_i\}$ and $\{w_j\}$ are the sets of the N most probable words for topics i and j , respectively, with $i \neq j$. This yields $K(K-1)/2$ pairwise overlap scores per model. The overlap score ranges from 0 to 1, where 0 indicates no shared top words (perfect uniqueness) and 1 indicates complete redundancy.

For our empirical analyses, we excluded stopwords to enhance the interpretability of the most probable words per topic. To this end, we carefully curated a list of 56 stopwords that we determined to be non-contributory to our topic analysis. After removing these stopwords, we extracted the 20 most likely words per topic, which served as the basis for subsequent analyses of topic uniqueness.

This decision is motivated by several considerations. As discussed above, topics frequently contain high-frequency words that occur across a wide range of contexts (e.g., "a", "be", "do", "it"). While these terms are legitimate outputs of topic models, they contribute little to semantic interpretation. Crucially, their presence inflates topic overlap scores, as these words are likely to appear among the top-ranked terms in multiple topics. Moreover, differences in overlap across models could be driven by models varying ability to distribute stopwords across topics. Excluding stopwords thus aligns the analysis with how practitioners and applied researchers typically interpret topic models: by focusing on the semantically meaningful top words, with non-informative terms filtered out. Nonetheless, to ensure that this procedure did not introduce bias, we replicated the topic uniqueness analysis without removing stopwords and also varied the number of top words used for overlap calculation. These robustness checks confirmed that the relative model performance remained stable, indicating that our main findings are not an artifact of stopword exclusion.

To quantify topic uniqueness, we applied the methodology proposed by Büschken & Allenby (2020), computing pairwise topic overlap using the 20 most likely terms per topic ($N = 20$). This yields $K(K - 1)/2$ pairwise overlap scores per model. The mean model-level overlap scores for each model and dataset are presented in Table 4.5. For the Luxury Hotels and Camping Tents datasets, labeled UD-LDA achieves the lowest mean topic overlap, indicating the most unique topics among the models. In the American Restaurants dataset, classical LDA yields the lowest mean overlap, with labeled UD-LDA identifying the second lowest overlap. However, aggregated overlap scores can obscure within-model variation. Some topics may be highly distinct, while others show considerable overlap. Capturing this variation is essential for understanding how consistently a model generates well-separated topics for a given dataset.

	Luxury Hotels	Camping Tents	American Restaurants
LDA	0.152	0.212	0.040
SC-LDA	0.244	0.242	0.167
AT-LDA	0.184	0.074	0.251
unlabeled UD-LDA	0.123	0.077	0.096
labeled UD-LDA	0.076	0.070	0.057

Table 4.5: Model level topic overlap given model and dataset, based on $N = 20$ most likely terms per topic (excluding stopwords).

For this reason, Büschken & Allenby (2020) do not rely solely on mean overlap scores. Instead, they analyze the full distribution of the $K(K - 1)/2$ pairwise topic overlap scores. This approach provides a more nuanced view of model behavior. A model with consistently unique topics will produce a right-skewed distribution with a mode near zero, indicating that most topic pairs share few or no terms. Figure 4.5 displays the distributions of pairwise overlap scores for all models and datasets.

The distributions in Figure 4.5 reveal two important insights. First, while no single model dominates across all datasets, labeled UD-LDA shows a strong and consistent tendency to generate more unique topics than other models that incorporate structural features. In the Luxury Hotels and Camping Tents datasets, the mode of the overlap distribution for labeled UD-LDA is visibly closer to zero, indicating a higher concentration of minimally overlapping topic pairs. Even for the American Restaurants dataset, where classical LDA performs slightly better, labeled UD-LDA still closely follows and outperforms the other structurally informed benchmarks.

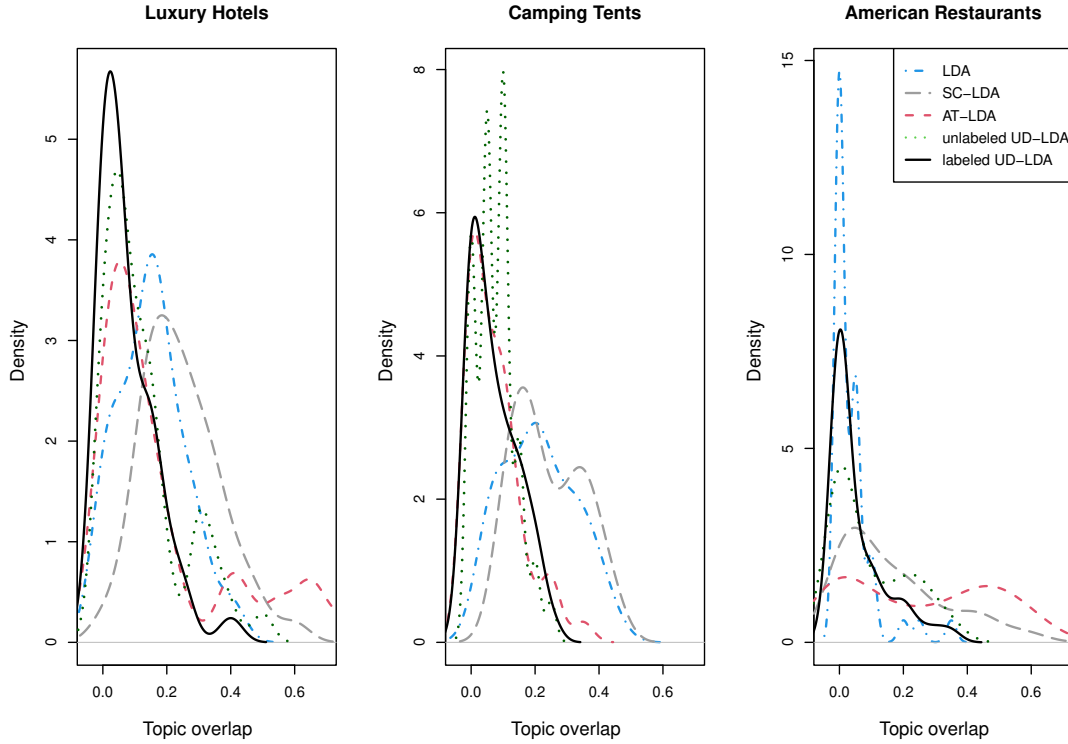


Figure 4.5: Distribution of topic overlap given model and dataset ($N = 20$, excluding stopwords).

Second, the distributions suggest that labeled UD-LDA produces fewer highly redundant topics. This pattern is most evident in the Luxury Hotels and Camping Tents datasets, where the density in the higher overlap range is comparatively lower for labeled UD-LDA. For the American Restaurants dataset, however, classical LDA again performs best in minimizing redundancy among topic pairs.

One possible explanation for these results is document length. The American Restaurants dataset contains significantly longer reviews, with an average of 190 words, whereas reviews in the Luxury Hotels and Camping Tents datasets average 53 and 88 words, respectively. Longer reviews offer richer word co-occurrence information at the document level, which may benefit classical LDA. In contrast, shorter reviews provide less such information, potentially increasing the value of structural text features for modeling topic distinctions. These hypotheses point to potential boundary conditions under which syntactic information enhances topic modeling performance. Future research is needed to examine these conditions more systematically.

Overall, we observe a consistent tendency for labeled UD-LDA to identify more unique topics, particularly in datasets with shorter documents. Importantly, our results indicate that the

improved model fit achieved by labeled UD-LDA does not come at the cost of more redundant topics. This reinforces the value of labeled UD structures in enhancing both model fit and topic uniqueness.

4.2.5 Topic coherence

Coherence is understood as an indicator of within-topic quality. In its literal sense, coherence refers to “*the situation in which parts of something fit together in a natural or reasonable way*”.³ In the context of topic modeling, this implies that humans should perceive the N most probable words within a topic as semantically related or fitting together in a natural or reasonable way. It is important to emphasize that coherence is not an intrinsic property of a topic model, but rather an empirical outcome that depends on the data. Whether a topic appears coherent hinges on the presence of meaningful word patterns in the dataset, which allow for interpretable top-word groupings. Nevertheless, comparing coherence scores across models applied to the same dataset can offer insights into their relative ability to produce semantically meaningful topics. In this way, coherence serves as a complementary performance criterion that helps determine whether improvements in model fit come at the cost of reduced interpretability.

The main challenge lies in **quantifying human-perceived coherence**. Manual evaluation across models, topics, and datasets is rarely feasible. To address this, various heuristic metrics have been developed and validated against human judgments (Mimno et al. 2011, Newman et al. 2010, Bouma 2009, Aletras & Stevenson 2013, O’callaghan et al. 2015). Appendix C provides a detailed overview of these measures and a systematic explanation of their components. To date, no single coherence measure has achieved universal acceptance. Different metrics capture different facets of coherence and may inherently favor certain model architectures. As a result, metric selection should be both theory-driven and transparently reported.

In this study, we adopt Word2Vec coherence as the evaluation metric. This decision is motivated by two key considerations (for details see Appendix C, Table C.1, second column):

1. **direct vs. indirect** confirmation measures
2. **intrinsic vs. extrinsic** reference corpus

³<https://dictionary.cambridge.org/us/dictionary/english/coherence>

Direct vs. indirect confirmation measures: Coherence measures assess the interpretability of a topic by analyzing its N most probable words and computing pairwise confirmation scores. A **direct confirmation measure** is based on raw word co-occurrence statistics. For example, the LCP measure (Mimno et al. 2011) evaluates how often two words appear together in the same documents, normalized by the frequency of the conditioning word. This method assumes that high co-occurrence frequency is indicative of semantic relatedness. By contrast, an **indirect confirmation measure** assesses coherence using vector-based word similarity. O’callaghan et al. (2015), for instance, propose computing cosine similarity between Word2Vec embeddings, which reflect the contextual usage of words. Words such as *football* and *soccer* might rarely co-occur, yet still receive a high similarity score they appear in similar contexts across the corpus.

This distinction is particularly relevant for the present work. All benchmark models considered in this study – except for classical LDA – integrate some notion of local context, whether based on word order, sentence boundaries, or syntactic dependencies. As such, evaluating coherence using document-level co-occurrence may systematically underestimate the performance of these more structure-sensitive models. In addition, our datasets consist of user-generated content, where individuals have considerable freedom in expressing ideas. People often express similar ideas using different words. While count-based measures may fail to recognize such variation, embedding-based metrics are designed to detect semantic similarity despite lexical diversity. Indirect, embedding-based coherence measures thus align more closely with both the model architectures under evaluation and the characteristics of the text data they are applied to.

Intrinsic vs. extrinsic reference corpus: Another key consideration is whether coherence should be evaluated using an intrinsic or extrinsic reference corpus. An extrinsic approach relies on an external, often large-scale corpus – such as Wikipedia – to compute word probabilities or word vectors, rather than using the dataset to which the topic model was applied. For example, Newman et al. 2010 argue that intrinsic methods risk reinforcing dataset-specific noise. However, their study dealt with large, heterogeneous corpora across broad domains. In contrast, marketing applications typically involve narrow, domain-specific data where such “noise” may actually reflect important characteristics of the dataset that merit closer investigation. We therefore use

an intrinsic Word2Vec coherence measure, computing confirmation scores on the same corpus to better reflect dataset-specific semantics.

Consistent with the approach used in the topic uniqueness analysis, we excluded stopwords from the list of top words in the coherence evaluation. Specifically, we applied the same list of 56 stopwords that were deemed non-informative for interpreting topic content. This ensures that both coherence and uniqueness analyses rely on a shared basis for evaluating top-word quality. The exclusion of stopwords is motivated by the observation that they offer little value for assessing within-topic coherence. Including stopwords among the top N most probable terms can lead to paradoxical results, particularly under co-occurrence-based metrics such as PMI (Fan et al. 2019). Because stopwords appear frequently and broadly across documents, they tend to co-occur with one another regardless of topic, often leading to artificially high coherence scores. However, such inflation does not reflect genuine semantic coherence or topic interpretability. Topic coherence is computed using the $N = 20$ most probable words per topic, excluding stopwords. We also replicated the analysis across alternative values of N and with stopwords included; the relative performance of the models remained stable across these conditions.

We generate word embeddings separately for each of the three datasets using the Word2Vec implementation in R (Mikolov et al. 2013). This dataset-specific approach ensures that the resulting word vectors reflect the distinct contextual usage of terms within each domain. For example, the semantic representation of the word *staff* may differ between Luxury Hotel and American Restaurant reviews, given its use in different contexts across these corpora. We train a Skip-gram model with a context window of ± 4 words and set the embedding dimensionality to $\text{dim} = 30$. The choice of parameters was guided by the characteristics of our datasets, which contain relatively few and short documents. To validate embedding quality, we manually inspected the most similar terms for selected seed words and confirmed that the resulting similarity structure was plausible.

Following O’callaghan et al. (2015), we compute the Topic-Level Word2Vec Coherence score (TC-W2V) as the average cosine similarity among all possible word pairs in the top N terms:

$$\text{TC-W2V} = \frac{1}{\binom{N}{2}} \sum_{j=2}^N \sum_{i=1}^{j-1} \text{sim}_{\cos}(\vec{v}_{w2v}^{(w_j)}, \vec{v}_{w2v}^{(w_i)}) \quad (4.2)$$

This yields one coherence score per topic. We calculate TC-W2V for all ten topics inferred by each model and aggregate these values to obtain a model-level mean coherence score for each dataset. Higher values indicate stronger semantic similarity among top-ranked topic words. The resulting scores are reported in Table 4.6.

	Luxury Hotels	Camping Tents	American Restaurants
LDA	0.284	0.274	0.211
SC-LDA	0.272	0.293	0.279
AT-LDA	0.259	0.336	0.207
unlabeled UD-LDA	0.278	0.331	0.234
labeled UD-LDA	0.292	0.311	0.250

Table 4.6: Mean model-level Word2Vec coherence (Skipgram Window Size $\pm 4L$), based on $N = 20$ most likely terms per topic (excluding stopwords).

We observe that aggregated Word2Vec coherence scores do not vary substantially across models within each dataset. In the Luxury Hotels dataset, labeled UD-LDA achieves the highest mean coherence score, closely followed by classical LDA. In the Camping Tents dataset, AT-LDA ranks highest, with unlabeled and labeled UD-LDA occupying the second and third positions, respectively. In the American Restaurants dataset, SC-LDA performs best, with labeled UD-LDA ranking second. While the relative ranking varies by dataset, these results do not indicate that any single model consistently outperforms others across all datasets. Labeled UD-LDA demonstrates robust performance across datasets, ranking among the top three models in all cases.

The full distributions of Word2Vec coherence scores further support this interpretation (Figure 4.6). While central tendencies are broadly similar across models, labeled UD-LDA shows a slight skew toward higher coherence values in the Luxury Hotels dataset. Similar patterns are observed for AT-LDA in the Camping Tents dataset and SC-LDA in the American Restaurants dataset. Taken together, these findings suggest that labeled UD-LDA performs competitively in terms of topic coherence. Importantly, it does so despite introducing structural constraints absent from standard LDA.

In response to the first research question – whether incorporating UD-based syntactic depen-

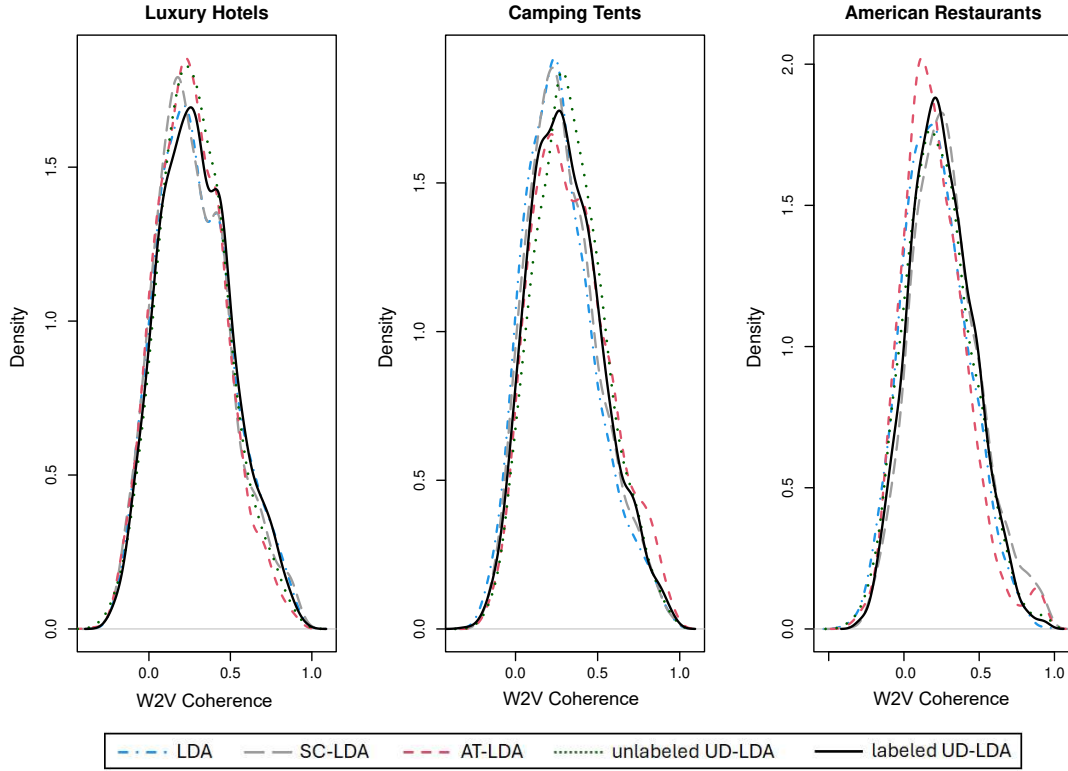


Figure 4.6: Distribution of w2v coherence given model and dataset ($N = 20$, excluding stopwords).

dencies enhances the representation of latent topic structure – we find strong empirical support for the proposed labeled UD-LDA model. Despite its structural constraints and reduced degrees of freedom, labeled UD-LDA achieves a favorable overall balance. It outperforms benchmark models in terms of in-sample model fit, produces more distinct (i.e., less overlapping) topics – indicating greater topic uniqueness – and yields competitive coherence scores. Crucially, these results should be interpreted in combination: the model’s ability to identify distinct topics without sacrificing coherence is a particularly desirable property. This balance suggests that improvements in model fit and topic distinctiveness do not come at the expense of interpretability. Lastly, we note that coherence was evaluated using automated metrics; future research may complement these findings with human assessments.

4.3 Explaining topic propagation through syntactic structure

This chapter investigates the upper level of the labeled UD-LDA model architecture. More specifically, it focuses on the mechanism of topic propagation, as captured by the latent indicator

variable ζ_{dsn} . This binary variable determines whether a given word inherits its topic assignment from its syntactic head or samples a new topic from the document-level topic distribution. The probability of topic propagation, denoted ψ_{dsn} , is modeled using a binary regression framework. It is a function of word-level covariates x_{dsn} and corpus-level regression coefficients β . The covariates include both the topic assignment of the syntactic head, z_{dsh_n} , and the type of syntactic dependency relation $dep_type(w_{dsh_n}, w_{dsn})$.

A detailed analysis of this model component provides insights into the second research question of this study, **RQ2**: *Does the type of syntactic dependency, as indicated by UD labels, contribute to explaining the probability of topic propagation along the edges of dependency trees?*

To address this question, the remainder of this chapter is structured as follows. Subsection 4.3.1 presents the regression results and identifies syntactic relations that significantly influence topic propagation probability. Subsection 4.3.2 provides qualitative examples and dependency-based visualizations to illustrate how syntactic structure shapes topic propagation in real sentences. Subsection 4.3.3 highlights the consistency of topic propagation patterns across datasets and discusses implications for both linguistic theory and model interpretability.

4.3.1 Empirical effects of dependency relation types

To assess the empirical influence of dependency relation types on topic propagation, we examine the corpus-level regression coefficients β_{dep_type} , which estimate the effect of each Universal Dependency relation type on the probability of topic propagation. Figures 4.7 and 4.8 display the posterior distributions of the MCMC draws for these β coefficients, based on the Luxury Hotels and Camping Tents datasets, respectively. The dependency types are ordered from strongest positive to strongest negative effects.

Coefficients that are not significantly different from zero suggest that the corresponding syntactic relation does not meaningfully affect topic propagation. This implies that syntax alone may be insufficient to explain topic transitions for these types. In contrast, coefficients that are significantly positive or negative indicate that the respective dependency relation systematically increases or decreases the probability of topic propagation. The results show that most syntactic

dependency types have a statistically significant effect on topic propagation probability. That is, the β coefficients for the majority of relation types is significantly different from zero, either positively or negatively. This supports the interpretation that syntactic structure systematically shapes the thematic relationship between syntactic heads and their dependents.

Importantly, we observe a high degree of consistency across datasets for relations that exhibit strong positive or negative effects. Dependency relations that substantially increase the likelihood of topic propagation in both datasets include “det” (determiner), “amod” (adjectival modifier), “advmod” (adverbial modifier), “cop” (copula), and “aux” (auxiliary). These results suggest that adjectives modifying nouns, adverbs modifying verbs, and function words such as determiners and auxiliaries tend to inherit the topic from their syntactic heads rather than introducing new topics. This aligns with linguistic expectations: modifiers serve to refine or contextualize the meaning of their heads rather than shift the topic, while function words are generally non-autosemantic and thus adopt the topic of their head.

Conversely, dependency relations associated with decreased topic propagation include “obj” (object), “obl” (oblique modifier), “acl” (adnominal clause), “acl:relcl” (relative clause), and “ccomp” (clausal complement). These constructions typically introduce clauses or nominals to a sentence. Clausal constructions represent events, actions or states, while nominal constructions introduce entities involved in those. The significantly negative β coefficients for these relations indicate a reduced probability of topic inheritance, reflecting their ability in introducing autosemantic content into the sentence.⁴

⁴Note that we also estimated a β coefficient for the “root” relation. However, deterministically set $\zeta_{n=root} = 0$, ensuring that root words never inherit topics from an undefined head (see Equation 3.5).

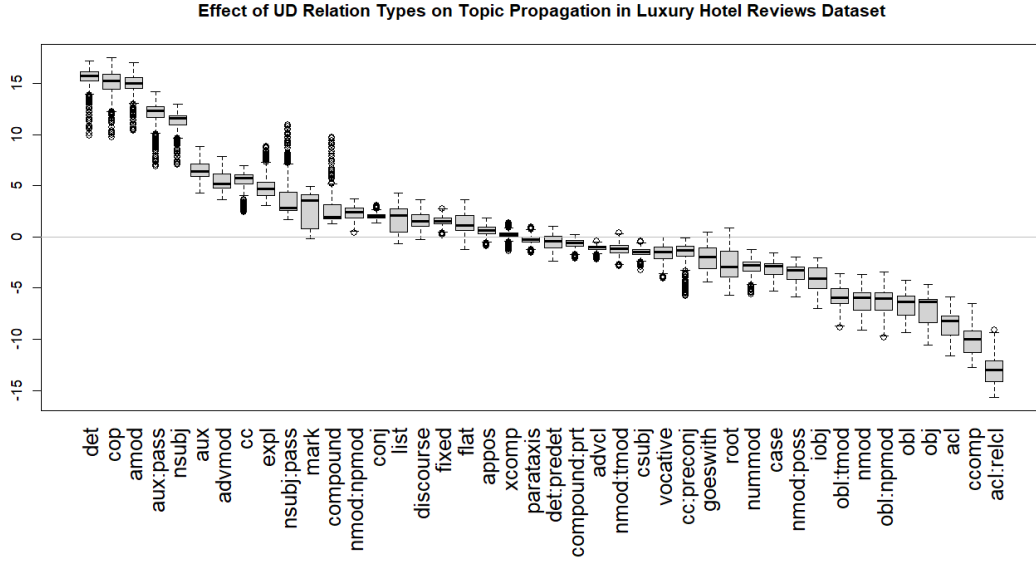


Figure 4.7: Posterior distribution of MCMC draws for β -coefficients, estimating the effect of Universal Dependency relationship types on topic propagation probability in Luxury Hotel reviews dataset.

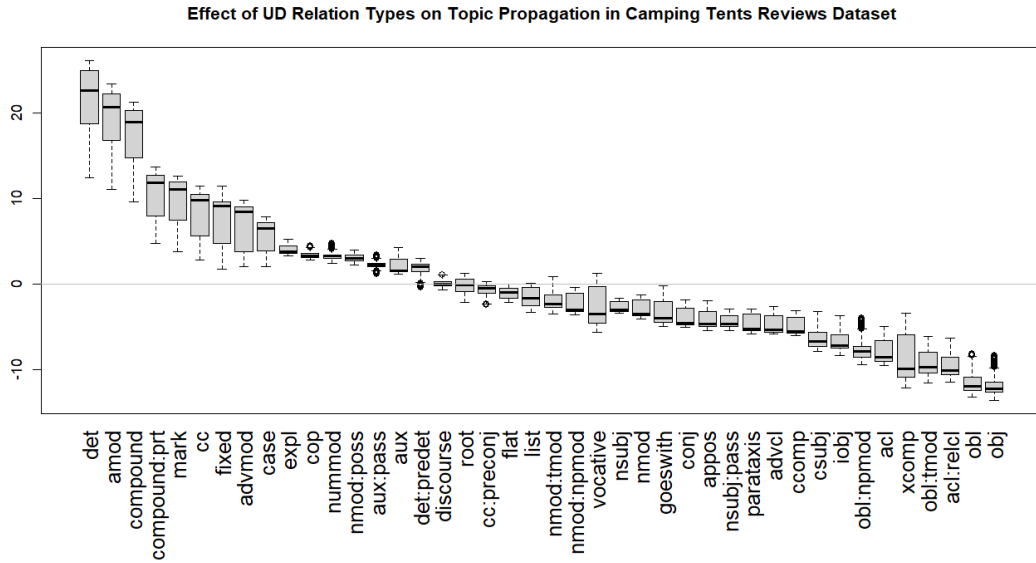


Figure 4.8: Posterior distribution of MCMC draws for β -coefficients, estimating the effect of Universal Dependency relationship types on topic propagation probability in Camping Tents reviews dataset.

4.3.2 Dependency-based visualization of topic propagation

While the estimated β coefficients provide a linguistically motivated explanation of how syntactic relations influence topic propagation, these insights remain somewhat abstract at the aggregate level. The effect of the hierarchical regression becomes more tangible when explored through concrete sentence-level examples. To illustrate this, we visualize the latent topic assignments in selected reviews from our datasets. Figures 4.9 and 4.10 present color-coded sentences from the Luxury Hotels and Camping Tents datasets, respectively. Each word is colored according to its topic assignment in the final iteration of MCMC sampling. Words that share the same color (within the same model and dataset) received the same topic assignment.

These examples reveal a pronounced contrast between standard LDA and the labeled UD-LDA model. In the standard LDA results (left-hand side), topic assignments reflect the bag-of-words assumption, under which words are treated as conditionally independent. As a result, changes in topic assignments often appear abrupt – even between syntactically or semantically linked expressions. For instance, phrases such as “Times Square” or “New York City” are not consistently assigned to the same topic. Similarly, in the tent review, numerical expressions like “\$ 99” or “20 minutes” are split across topics. Although standard LDA may occasionally assign adjacent words to the same topic (e.g., “New” and “York” in the first review), such outcomes are incidental rather than structurally informed.

In contrast, the labeled UD-LDA results (right-hand side) display visibly more coherent groupings of topic assignments. Expressions such as “Times Square”, “New York City”, and number-unit pairs like “\$ 99” and “20 minutes” are more consistently assigned to a single topic. This alignment of topic assignments is driven by the model’s hierarchical regression on UD relations, where certain dependency types – such as modifiers and compounds – systematically promote topic propagation. As a result, the labeled UD-LDA model enforces topical continuity in places where syntactic links imply semantic interdependence. The resulting topic assignments are not only syntactically grounded but also better aligned with human intuitions regarding which words “belong together” thematically.

LDA	labeled UD-LDA
The hotel was great. Very clean rooms and a wonderful view of Times Square. The staff was very friendly and helpful. I would definitely stay again next time I visit New York City. This is a great value. My only complaint is that we got a lot of street traffic noise off and on until about 2 in the morning. But hey this is the city that never sleeps so what do you expect. We had a brand new flat screen hdtv. The room was a decent size considering this is New York City. Bathroom and shower both very nice. The building has retained its historic charm look and feel.	The hotel was great. Very clean rooms and a wonderful view of Times Square. The staff was very friendly and helpful. I would definitely stay again next time I visit New York City. This is a great value. My only complaint is that we got a lot of street traffic noise off and on until about 2 in the morning. But hey this is the city that never sleeps so what do you expect. We had a brand new flat screen hdtv. The room was a decent size considering this is New York City. Bathroom and shower both very nice. The building has retained its historic charm look and feel.

Figure 4.9: Color-coded example reviews from the Luxury Hotels dataset. Each color represents a unique latent topic assignment based on the final state of the MCMC sampling. The left column shows standard LDA results, while the right column shows topic assignments from the labeled UD-LDA model.

LDA	labeled UD-LDA
This tent is huge. Took 15 minutes to put up the first time and another 15 minutes to pack it up. The power corners really add stability and the rear mesh vent works well for airflow. We purchased this tent for \$ 99 and free shipping. We set it up in our backyard it took 15 minutes for the main part the rainfly part was a little tricky. I would say 20 minutes for that but it may be easier the next time.	This tent is huge. Took 15 minutes to put up the first time and another 15 minutes to pack it up. The power corners really add stability and the rear mesh vent works well for airflow. We purchased this tent for \$ 99 and free shipping. We set it up in our backyard it took 15 minutes for the main part the rainfly part was a little tricky. I would say 20 minutes for that but it may be easier the next time.

Figure 4.10: Color-coded example reviews from the Camping Tents dataset. Each color represents a unique latent topic assignment based on the final state of the MCMC sampling. The left column shows standard LDA results, while the right column shows topic assignments from the labeled UD-LDA model.

We can further illuminate the underlying topic propagation mechanism by visualizing how topics propagate through the syntactic structure of sentences. While the final topic assignments highlight patterns of topical alignment, they do not reveal the underlying mechanism of topic propagation – that is, which topic assignments are inherited from syntactic heads and which are drawn independently from the document-level topic distribution θ_d .

Figures 4.11 and 4.12 present example sentences annotated with their UD parse structures, which are treated as observed in the labeled UD-LDA. In these visualizations, black arcs indicate the absence of topic propagation (i.e., the topic is drawn from θ_d). Colored arcs indicate topic propagation from the syntactic head to the dependent (i.e., $\zeta = 1$). Each color corresponds to a distinct topic.

These visualizations provide a more granular view of how the labeled UD-LDA model leverages syntactic structure to propagate topics. For instance, in the sentence “*Very clean rooms and a wonderful view of Times Square*”, the noun “rooms” (root) receives its topic directly from θ_d . This topic is then propagated to its modifiers “clean” (amod) and “very” (advmod), – both of which are dependency types with strongly positive β coefficients in the model.

A similar propagation mechanism is evident in compound constructions such as “*Times Square*”, “*New York City*”, “*power corners*”, and “*mesh vent*”. In each case, the compound relation promotes topic propagation, even across multiple words or against the typical reading direction, due to its significantly positive effect on topic propagation probability.

At the same time, numerous breaks in topic propagation are observed, particularly for dependency types such as objects (obj), oblique modifiers (obl), and conjuncts (conj), which are associated with low β coefficients. These syntactic roles often introduce new entities or events and thus serve as natural boundaries in topic flow, especially in more complex sentences involving multiple predicates and (in)direct arguments. This pattern is clearly observable in the final sentence of Figure 4.12, where two predicates – “add” and “works” – govern separate syntactic units: “*The power corners really add stability*” and “*the rear mesh vent works well for airflow*”. While being connected via a conjunction (conj), each predicate introduces its own action. The absence of topic propagation between the two predicates is indicated by a black arc, reflecting that topic propagation does not occur across this syntactic boundary. Additionally,

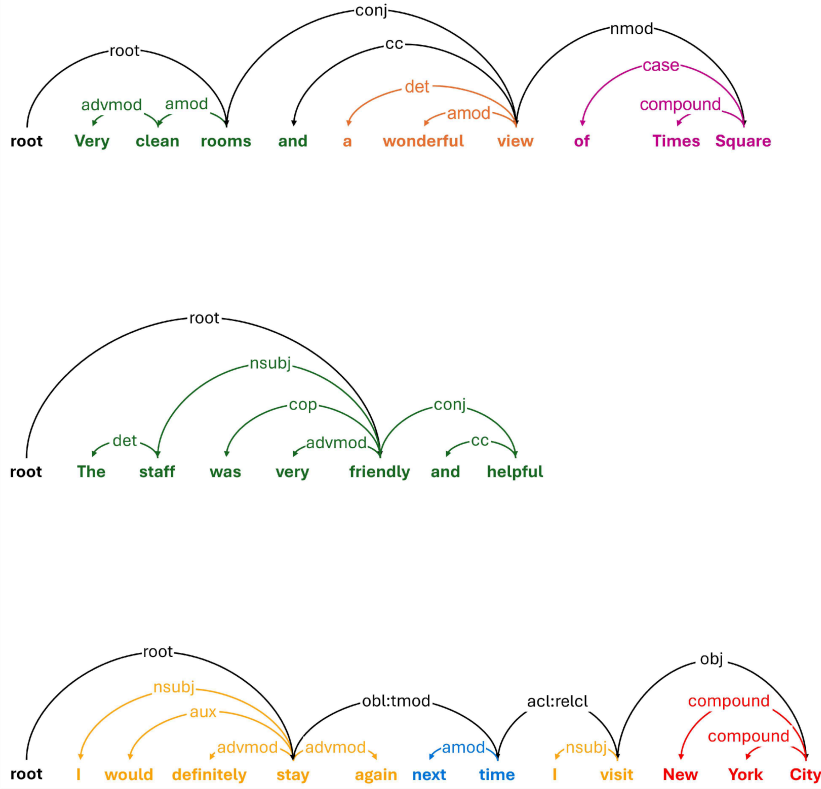


Figure 4.11: Dependency-based visualization of topic propagation in sentences from the Luxury Hotels dataset. Arcs and labels represent syntactic dependencies from UD parses. Colored arcs indicate topic propagation between head and dependent (i.e., $\zeta = 1$), while black arcs represent absence of topic propagation (i.e., $\zeta = 0$).

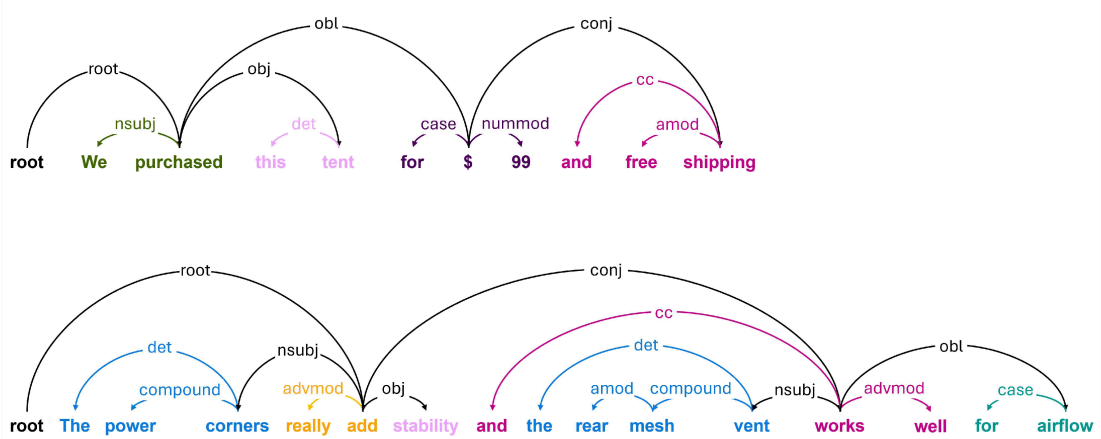


Figure 4.12: Dependency-based visualization of topic propagation in sentences from the Camping Tents dataset. Arcs and labels represent syntactic dependencies from UD parses. Colored arcs indicate topic propagation between head and dependent (i.e., $\zeta = 1$), while black arcs represent absence of topic propagation (i.e., $\zeta = 0$).

topic propagation is suppressed between each predicate and its core arguments. For instance, “add” does not propagate its topic to “corners” (nsubj) or “stability” (obj). Similarly, “works” does not transfer its topic to “vent” (nsubj) or “airflow” (obl). This lack of topic propagation aligns with linguistic theory, which posits that core syntactic UD relations often involve words with autosemantic properties. Such words that introduce distinct topical content therefore receive independent topic assignments in the labeled UD-LDA model.

Together, these dependency-based visualizations offer a mechanistic view of how the labeled UD-LDA model integrates syntactic structure with probabilistic topic inference. Conditioning topic propagation on UD relations leads to more consistent and semantically meaningful topic assignments. Through relation-specific β coefficients, the model promotes topic propagation for modifiers, compounds and function words – relations by which words are closely linked and jointly refer to a concept, whether an entity or an event. Conversely, propagation is suppressed for arguments and clausal relations, which typically introduce new topical content. This reflects a core linguistic insight: syntactic form signals thematic relatedness

4.3.3 Topic propagation patterns across datasets

The preceding analyses have demonstrated that UD relations provide meaningful structural cues for modeling topic propagation, by conditioning the probability of topic inheritance on syntactic relation types. So far, these effects have been interpreted through the lens of the regression coefficients β_{dep_type} , which capture how each dependency relation systematically affects the probability of topic propagation. To complement this analysis with a more intuitive perspective, we now directly examine the augmented binary topic propagation variable ζ_{dsn} . This variable indicates whether a word inherits its topic from its syntactic head ($\zeta_{dsn} = 1$) or receives an independent topic assignment from the document-level distribution ($\zeta_{dsn} = 0$). The goal is to identify which dependency relations systematically promote or inhibit topic propagation, and whether these effects generalize across datasets.

We summarize topic propagation behavior at the level of dependency relations by computing the frequency for each relation type r as follows:

$$f_r = \frac{\sum \zeta_{dsn}}{N_r} \quad \text{for all } (d, s, n) \text{ where } \text{dep_type}(w_{dsn}) = r$$

Here, the numerator counts the number of topic inheritance events for relation type r , while N_r denotes the total number of instances of that relation in the corpus. The resulting proportion lies in the interval $[0, 1]$, where 1 indicates that all dependents of type r inherit the topic from their head and 0 indicates that none do. This measure offers an interpretable, relation-level summary of topic propagation behavior and serves as the basis for cross-dataset comparison.

Before delving into relation-specific findings, we first assess whether topic propagation behavior is consistent across datasets. For each dataset, we calculate the empirical topic propagation frequency for every dependency relation type – resulting in a vector of length J , where J denotes the number of distinct UD relation types observed. We then compute Pearson correlation coefficients between these vectors to evaluate how similarly dependency types influence topic propagation across datasets. Table 4.7 presents the resulting correlation matrix. The correlation between Luxury Hotels and Camping Tents is 0.7839, and similarly high correlations are found between Luxury Hotels and American Restaurants (0.8198) and Camping Tents and American Restaurants (0.6856). These results provide a first indication of cross-domain consistency: High correlations suggest that the influence of syntactic relation types on topic propagation is largely stable across datasets.

	Luxury Hotels	Camping Tents	American Restaurants
Luxury Hotels	1	-	-
Camping Tents	0.7839	1	-
American Restaurants	0.8198	0.6856	1

Table 4.7: Pearson correlation coefficients of relation-level topic propagation frequencies across the three datasets.

This consistency becomes especially evident when visualizing the propagation frequencies across datasets. Figure 4.13 plots relation-level topic propagation frequencies for the Camping Tents dataset (x-axis) against those for the Luxury Hotels dataset (y-axis). Each point represents a distinct UD relation type, with coordinates reflecting the proportion of topic propagation instances for that relation in each dataset.

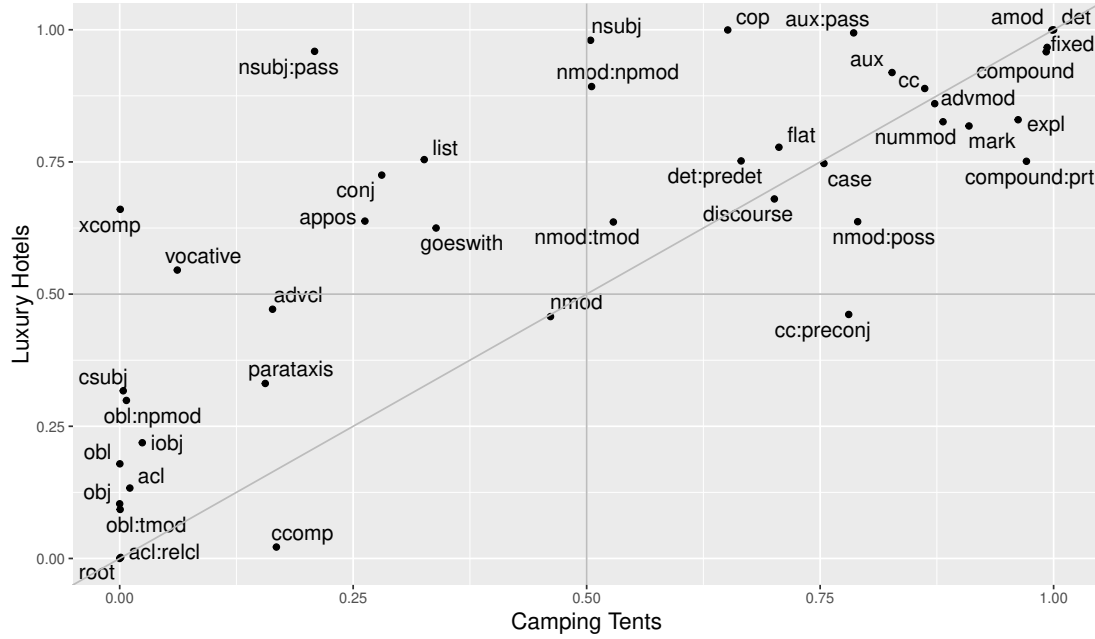


Figure 4.13: Relation-level topic propagation frequencies in the Luxury Hotels and Camping Tents datasets.

A clear pattern emerges: most relation types cluster closely along the diagonal, indicating strong agreement in propagation behavior between the two domains (Pearson correlation = 0.7839). Relations with consistently strong positive or negative associations with topic propagation appear in the upper-right or lower-left quadrants, respectively – signaling that their influence is both systematic and stable across datasets.

Among the relations with consistently high topic propagation frequencies (> 0.7), we find *amod*, *advmod*, *compound*, *fixed*, and *flat*. These relations commonly appear between words that form nominal or clausal constructions and support thematic continuity among closely linked word pairs. Likewise, dependency types that connect function words to content heads such as – *det*, *cop*, *cc*, *mark*, *case*, *expl*, or *aux* – also appear in the upper-right quadrant, typically exhibiting propagation frequencies above 0.6. This reflects the tendency of function words to inherit the topic from their syntactic heads.

In contrast, the bottom-left quadrant features relations such as *obj*, *obl*, *acl*, and *advcl*, which typically link distinct syntactic constructions. These relations are associated with low propagation frequencies, supporting the idea that such links connect syntactic units with independent thematic content.

Taken together, these patterns underscore a high degree of structural regularity in topic propagation across datasets. They reinforce the conclusion that syntactic relations, as captured by Universal Dependencies, provide linguistically meaningful signals for modeling topic interdependencies in text.

We found that relation-level topic propagation frequencies are highly consistent across datasets. This motivates a closer examination of those dependency types that consistently exhibit high or low topic propagation tendencies across all datasets.

Table 4.8 summarizes these patterns for all three datasets – Luxury Hotels, Camping Tents, and American Restaurants. The table is organized according to the UD typology introduced in Section 2.1.3.2, grouping relation types by syntactic function. To aid interpretation, it uses color coding to highlight consistent cross-dataset behavior: green cells indicate relations with high propagation frequencies (> 0.5) across all datasets, while red cells denote relations with consistently low propagation frequencies (< 0.5). White cells represent relation types with inconsistent patterns, where topic propagation behavior varies across datasets, precluding a clear interpretation.

High topic propagation frequencies: Two broad syntactic configurations are associated with consistently high topic propagation frequencies:

1. relations **within nominal or clausal constructions**

- (a) **modifier relations**

- (b) **special syntactic constructions**

2. **function words** attached to content heads

Modifier relations – including *advmod*, *amod*, *nummod*, *nmod:npm*, *nmod:poss*, and *nmod:tmod* – tend to inherit the topic of their syntactic head. This pattern aligns with the content-head principle in UD, which posits that heads of core relations are autosemantic – they retain meaning independently of context – while modifiers serve to refine or contextualize the meaning of those heads (De Marneffe et al. 2021). For example, in the adjectival modifier (*amod*) relation : (e.g., “good” $\leftarrow$ “care”), the adjective “good” sharpens the meaning of the noun “care” without introducing new topical content. The labeled UD-LDA model captures this asymmetry

	Nominals	Clauses	Modifier words	Function Words
Clausal Core arguments	nsubj	csubj		
	nsubj:pass	ccomp		
	obj	xcomp		
	iobj			
Clausal Non-core dependents	obl	advcl	advmod	aux
	vocative		discourse	aux:pass
	expl			cop
	dislocated			mark
Nominal dependents	nmod	acl	amod	det
	nmod:npmode	acl:relcl		det:predet
	nmod:poss			case
	nmod:tmod			
	appos			
	nummod			
Coordination	Headless	Loose	Special	Other
conj	fixed	list	compound	root
cc	flat	parataxis	compound:prt	
			orphan	
			goeswith	
			reparandum	

Table 4.8: Colorcoded relation-level topic propagation frequencies across all three datasets examined (Luxury Hotels, Camping Tents, and American Restaurants). Green: Topic propagation frequencies >0.5 for all three datasets. Red: Topic propagation frequencies <0.5 for all three datasets.

by inferring high topic propagation probabilities from head to modifier, consistent with both linguistic theory and empirical behavior.

The same logic applies to **special syntactic constructions** like *compound* and *fixed*, which bind words into multi-word expressions. These include noun compounds (e.g., “counter space”, “New York City”) and fixed phrases (e.g., “less than”, “because of”). Topic propagation is intuitively expected in such constructions: the individual words are semantically interdependent and jointly form a unified concept. This is reinforced by cross-linguistic patterns – for example, in German, noun compounds like “counter space” are lexicalized into single words (“Arbeitsfläche”). Labeled UD-LDA captures this structural cohesion by promoting topic propagation within such constructions, in line with their syntactic and semantic unity.

The second major syntactic source of high topic propagation involves **function words**. Relations such as *det* (e.g., “the” $\leftarrow$ “hotel”), *case* (e.g., “for” $\leftarrow$ “me”), *cop* (e.g., “were” $\leftarrow$ “pleased”), *expl* (e.g., “there” or “it”), and *aux* (e.g., “could” $\leftarrow$ “go”) all exhibit high propagation frequencies across datasets. The role of function words in topic modeling has long been debated.

Function words have been excluded under the assumption that they lack topical content (Tirunillai & Tellis 2014, Martin & Johnson 2015), or treated in separate latent classes (Griffiths et al. 2004). The UD framework offers a more nuanced view: function words are structurally subordinate elements, typically low in the dependency tree and connected to semantically richer content heads (Nivre et al. 2016). Labeled UD-LDA leverages this structural asymmetry – function words almost never introduce new topics but instead adopt the topical context of their head. This pattern is confirmed consistently across all three datasets.

Low topic propagation frequencies: As noted earlier, dependency types that **link distinct syntactic units** tend to suppress topic propagation. These relations show a consistent tendency for the dependent to receive a new topic assignment rather than inherit one. We distinguish three major subtypes:

1. **nominal dependents** of clausal heads
2. **clausal dependents**, and
3. **loose constructions**.

Nominals dependents attached to clausal heads, such as (in)direct objects (*obj*) and oblique modifiers (*obl*) consistently exhibit low topic propagation frequencies. These relations typically introduce entities – such as people, objects, or locations – into a sentence. For example: “*We **saw** (head) a few Broadway **shows** (dependent)*”, where “shows” is attached to its clausal head “saw” via an object relation. Or: “*the main subway line **got** (head) us all over the **city** (dependent)*”, where “city” is linked to its clausal head “got” via a prepositional *obl* relation. In both cases, the dependent nouns refer to novel entities and represent new thematic content. These are core syntactic relations whose dependents are assumed to be autosemantic (De Marneffe et al. 2021). Their semantic independence explains the model’s preference for assigning them distinct topics.

Relations introducing **clausal dependents** such as *csubj*, *ccomp*, *acl*, and *acl:relcl* also show consistently low topic propagation. These dependencies introduce full clauses, each typically headed by a predicate and accompanied by its own arguments and modifiers (De Marneffe et al. 2021). Consider the sentence: “*As a lady who **went** (dependent) alone I **felt** (head) very secure at this hotel*”. The relative clause “*who went alone*” is linked to the main clause “*I felt...*” via the

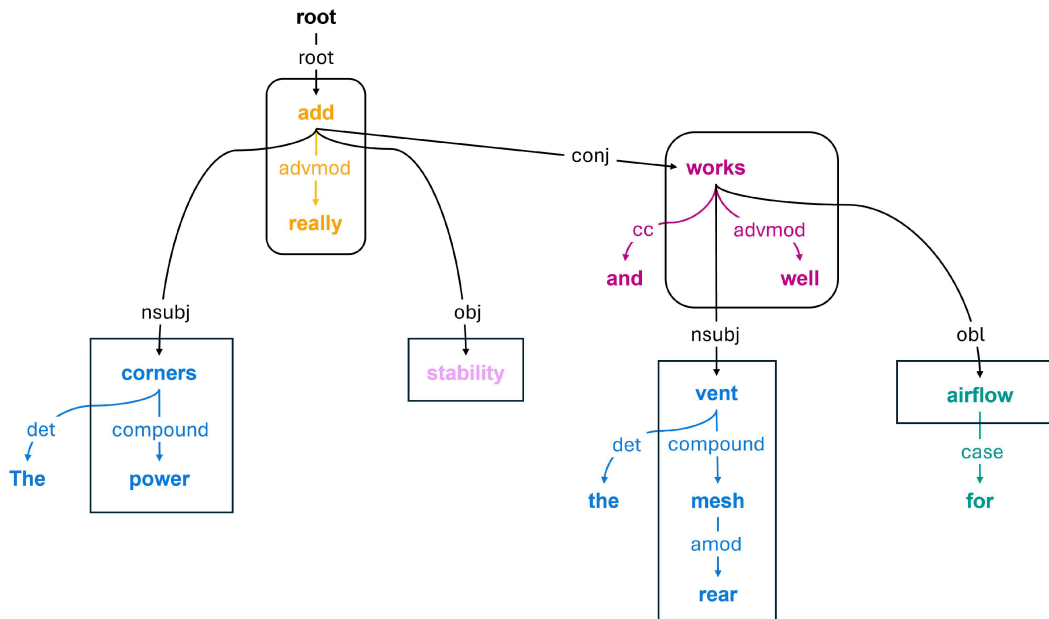


Figure 4.14: Visualization of relation-specific topic propagation in a UD-annotated sentence.

Color-coded arrows indicate topic propagation from head to dependent ($\zeta = 1$), while black arrows mark the absence of propagation ($\zeta = 0$). Rectangular and rounded boxes denote nominal and clausal constructions, respectively, highlighting distinct topic propagation behavior within and across syntactic units.

acl:relcl relation. Despite this syntactic link, the dependent clause expresses its own event. The labeled UD-LDA reflects this autonomy through low topic propagation probabilities.

Finally, **loosely connected syntactic constructions** – such as *parataxis* – exhibit low topic propagation. These relations connect independent clauses or nominals without subordination or coordination and are often signaled by punctuation (De Marneffe et al. 2021). For example, in the Luxury Hotels dataset: “... *they **said*** (head) *they would send an engineer; no one ever **came*** (dependent)”. The semicolon marks a boundary between syntactically loosely connected clauses. Consequently, the model assigns separate topics to the two events.

Figure 4.14 illustrates the three main trends in how syntactic relations systematically shape topic propagation probabilities, using an example sentence from the Camping Tents dataset in UD tree structure. Colored arrows indicate topic propagation between head and dependent ($\zeta = 1$), while black arrows mark the absence of topic propagation ($\zeta = 0$). Color-coded topic assignments reflect the final state of the MCMC chain. The topic assignments themselves correspond to the final state of the MCMC chain. Nominal constructions are enclosed in rectangular boxes,

while clausal constructions are shown with rounded rectangles, visually distinguishing the two foundational syntactic units in UD.

Three core patterns – previously identified in the empirical analysis – become visually apparent:

(1) Topic propagation is frequent **within** nominal and clausal constructions. Relations such as *amod*, *advmod* or *compound* link words that jointly refer to a single semantic unit. In nominals, these relations form cohesive entities (e.g., “*power corners*”, “*rear mesh vent*”); in clauses, they represent events (e.g., “*really add*”, “*works well*”). Accordingly, the topic assigned to the syntactic head is typically propagated to its dependents, resulting in uniform topic assignments. This is reflected in Figure 4.14 by the colored arrows within boxes.

(2) Topic propagation is less common **across** distinct constructions. Relations such as *nsubj*, *obj*, *obl*, and *conj* connect the heads of different nominal or clausal units. These syntactic links often introduce new thematic content, marking boundaries between topical segments. In Figure 4.14, these cross-unit connections are represented by black arrows between boxes, indicating the absence of topic propagation.

(3) Function words typically inherit the topic of their head. Syntactic relations that link function words to content heads – such as *det*, *case*, and *cc* – exhibit consistently high topic propagation. These words provide grammatical structure rather than introducing new semantic content. In Figure 4.14, these relations are represented by colored arrows connecting the function words to their heads, reinforcing their role as syntactically and thematically subordinate elements.

These findings demonstrate that topic propagation in labeled UD-LDA is governed by systematic syntactic patterns. The model regulates topic propagation based on grammatical function: Promoting propagation within cohesive syntactic units, and allowing for topic shifts across distinct syntactic units. Crucially, it is not merely the information about which words are connected via head-dependent relations that explains topic propagation, but the specific **type of syntactic relation** – highlighting grammar’s role as a structural signal for thematic organization.

Chapter 5

Discussion

This thesis integrates two complementary perspectives on statistical text analysis to enhance applied marketing research. In contexts where data availability is inherently limited due to the targeted nature of the analysis, it becomes essential to make efficient use of available textual information. Here, “efficiency” refers to using not only the words in text, but also the structure of language (i.e., *how* text is structured). This work builds on the idea that syntactic structure – in the form of Universal Dependencies (UD) – can provide valuable information to support topic inference.

Universal Dependencies encode word-level syntactic functions in the form of labeled head-dependent relations, which can be obtained using off-the-shelf dependency parsers. These labeled structures provide a linguistically principled basis for modeling topic propagation within sentences, as the labels reflect not only syntactic form but also semantic relationships between words. The labeled UD-LDA model developed in this thesis incorporates UD into topic modeling by conditioning topic propagation along the branches of dependency trees.

Labeled UD-LDA is the first topic model to conceptually integrate *labeled* syntactic dependencies with the modeling of topic propagation in text. This thesis shows that syntactic dependencies exhibit three central characteristics: they are predominantly local, systematically include non-adjacent relations, and frequently run against the reading direction, with dependents preceding their heads. These properties imply that the relevant context for topic interdependencies is not only short-ranged but also often extends beyond immediate neighbors and can

lie to the right of a token in the sequence. Conventional topic models that rely on sequential proximity or first-order assumptions cannot capture this complexity, while more recent distributional approaches approximate it only through large-scale, data-driven model architectures. UD, by contrast, offer a theory-informed representation that explicitly encodes both adjacency and directionality of syntactic relations. In this way, UD provides a transparent and linguistically grounded signal for topic inference, particularly valuable in targeted, data-scarce marketing applications.

The empirical application of labeled UD-LDA addressed the following research question:

RQ1: Does a topic model that conditions topic propagation on UD-based syntactic dependencies yield an enhanced representation of the latent topic structure in text datasets?

To answer this question, the labeled UD-LDA was benchmarked against several alternative models, including standard LDA, across three datasets from distinct marketing domains. The results demonstrated that labeled UD-LDA consistently outperformed all benchmarks in terms of in-sample model fit. Additionally, it exhibited improvements in topic uniqueness and maintained competitive performance in topic coherence. Importantly, these gains were achieved despite the model's more constrained structure, indicating that the incorporation of syntactic constraints does not compromise topic interpretability.

The second research question examined the syntactic foundations of topic propagation:

RQ2: Does the type of syntactic dependency, as indicated by UD labels, contribute to explaining the probability of topic propagation along the edges of dependency trees?

The hierarchical Bayesian framework of the labeled UD-LDA allows for a detailed investigation of how different dependency relations influence topic propagation. Two key findings emerged. First, syntactic relation types were found to have systematic and statistically significant effects on topic propagation. Relations within syntactic units – such as modifiers in nominals or clauses – and relations connecting function words to content heads, consistently increased the probability of topic propagation. In contrast, relations bridging syntactic units (e.g., subject,

object, oblique, and clause relations) tended to suppress topic propagation. These results support the hypothesis that grammatical structure plays a central role in shaping the flow of thematic content in text.

Second, the observed effects were highly consistent across datasets and aligned closely with linguistic theory. According to the UD framework, syntactic heads of core constructions are typically semantically autonomous and more likely to introduce new topical content, whereas dependents – particularly modifiers and function words – tend to refine or support the meaning of their heads. These theoretical expectations are reflected by results from our model.

Collectively, these findings demonstrate that syntactic structure – as formalized by the Universal Dependencies framework – constitutes a meaningful signal for modeling topic propagation in text. The labeled UD-LDA offers a transparent and linguistically grounded approach to integrating observed syntactic features into probabilistic topic models. This integration contributes to both empirical performance and theoretical interpretability, thereby bridging the domains of automated text processing and language research. In doing so, the model aligns with the grammar conception proposed by Futrell et al. (2020), which views grammar as a system that mediates between linguistic form and semantic content.

5.1 Main contributions

This thesis was positioned at the intersection of “how” and “what”: The structural analysis of language (via dependency parsing) and the content-based analysis of text (via topic modeling), integrated within a targeted marketing analysis framework. This interdisciplinary perspective yields contributions to three research domains:

Language Research: In the field of language research, this work highlights the practical relevance of Universal Dependencies in probabilistic topic modeling. As noted by De Marneffe et al. (2021), the UD framework continues to grow rapidly, supported by a large community and extensive multilingual treebanks. Despite its foundational role in syntactic annotation, the application of UD to downstream modeling tasks remains relatively underexplored. This thesis introduces a novel use case by demonstrating how UD structures can inform topic inference,

that is, how topics are propagated across syntactic dependencies in text. The labeled UD-LDA is the first probabilistic topic model that incorporates UD labels and empirically examines how syntactic relations shape topic propagation probabilities. This thesis offers empirical evidence that UD encodes not only grammatical hierarchy but also a latent thematic structure, thereby supporting theoretical claims about the autosemantic properties of syntactic heads.

Topic Modeling: From a topic modeling perspective, this work contributes a linguistically grounded framework for probabilistic modeling of high-level topics as summaries of (large) text corpora. Previous efforts have used observed structural cues such as word order or sentence boundaries to approximate thematic continuity. This work advances the field by leveraging labeled dependency trees to condition topic propagation. Empirical results show that different types of syntactic relations exert distinct effects on topic propagation and that these effects align with linguistic expectations. This suggests that topic propagation can be more effectively modeled when informed by syntactic structure, rather than heuristic cues.

Marketing Research: For marketing research in data-scarce environments, the labeled UD-LDA offers a transparent and data-efficient approach to text analysis. Dependency parses can be generated with minimal preprocessing effort using standard tools and provide a valuable structural layer to complement content-based analysis. Unlike black-box models, the labeled UD-LDA has fully interpretable assumptions and does not rely on pretraining over large, external corpora. This transparency allows each inferred topic to be directly linked to the observed words and their syntactic relations, facilitating validation and interpretation. Such methodological clarity is especially critical in applied contexts, where decisions may depend on the validity of thematic patterns inferred from text.

In summary, this thesis proposes an interdisciplinary modeling framework that leverages syntactic structure to enhance both the empirical quality and theoretical grounding of topic inference. By integrating linguistic insights into probabilistic modeling, the labeled UD-LDA offers a robust and interpretable tool for analyzing text in targeted marketing settings and beyond.

5.2 Limitations and future research directions

The empirical analyses revealed consistent patterns of topic propagation across the three datasets, supporting the hypothesis that syntactic structure offers a generalizable signal for modeling topic propagation in sentences. However, all datasets were limited to user-generated English-language customer reviews, which are relatively homogeneous in style and structure. Future research should examine whether similar patterns emerge in other textual domains (e.g., news articles, academic writing) and in other languages. The cross-linguistic design of the UD framework makes it well suited for such comparative investigations.

While the labeled UD-LDA achieved competitive results in topic coherence, it did not consistently outperform all benchmark models. In line with standard practice, the evaluation relied on analyzing the most probable words under each topic. Such top-word lists provide an intuitive and widely used way of assessing topic quality, but they inevitably shift the focus back to the (individual) word level. This risks overlooking the core contribution of a syntactically informed model, which lies precisely in capturing head-dependent relations and syntactic patterns of language use. More syntactically aware approaches to evaluation could better reflect the strengths of the model – for example, analyzing topic-specific phrases (e.g., Büschken & Allenby 2025) or topic-specific grammatical constructions. Developing such grammatically informed methods of topic evaluation remains an open avenue for future research.

It is important to note that the current model formulation represents only one possible instantiation of syntactically informed topic modeling. More importantly, this thesis demonstrates that incorporating labeled dependency trees into a probabilistic framework can meaningfully capture topic interdependencies and support linguistic interpretation. Nonetheless, alternative model architectures – such as those expanding the covariate space or incorporating interaction effects – may yield further gains.

A promising avenue for model refinement involves extending the regression component of the topic propagation mechanism. The binary topic propagation indicator ζ can be flexibly conditioned on any observable word-level features. The UD framework provides additional morphosyntactic information, such as POS tags, which may reveal interaction effects between syntactic function and word class. For instance, both proper nouns and pronouns can serve

as nominal subjects (*nsubj*), yet they differ in semantic specificity (De Marneffe et al. 2021): Proper nouns, which typically denote specific entities such as people, places, or concepts, are semantically rich and more likely to introduce new topical content – hence, they may exhibit lower topic propagation probabilities. Pronouns, by contrast, function as placeholders for previously mentioned nouns and are generally less autosemantic; as such, they may more likely inherit the topic of their syntactic head. Accounting for such interactions between syntactic function and word class may lead to a more nuanced understanding of topic propagation.

Another limitation is that the labeled UD-LDA model treats syntactic structure as observed, assuming that the dependency tree is accurately known. In reality, language generation is not a sequential process of first selecting syntactic structure and then populating it with content; rather, syntax and semantics co-develop during language production. Indeed, one might even argue that the process can be reversed: in practice, writers often begin with raw ideas or bullet points and only subsequently embed them into syntactically well-formed sentences. Future research could explore generative models in which syntactic structure and topic assignments are inferred jointly, better approximating the dynamics of natural language generation and offering new directions for theory-driven modeling.

Finally, it is important to acknowledge that the Universal Dependencies framework itself is a dynamic and evolving resource. Since 2015, new UD treebank versions have been released approximately every six months, often introducing changes to the inventory of dependency relations. For instance, as of version 2.15 (released November 2024), certain relation types used in this work have been deprecated or merged (e.g., *nmod:tmod* and *nmod:npmode* have been merged into *nmod:unmarked*; *obl:tmod* and *obl:npmode* have been merged into *obl:unmarked*).¹ These developments do not affect the validity of the labeled UD-LDA’s model architecture, which is agnostic to the specific set of dependency labels. Moreover, the deprecated labels were not among those found to have strong effects on topic propagation and were not central to the model’s explanatory contributions. Nevertheless, future work should consider how to accommodate evolving annotation schemes. Potential strategies may include collapsing subtypes into higher-level categories (e.g., mapping *nmod:xxx* to *nmod*) or conditioning on a limited set of

¹<https://universaldependencies.org/>

frequent relations that exert the strongest influence on topic propagation. Exploring these design choices would support the development of leaner models that remain effective as linguistic frameworks evolve.

Seen from this perspective, the reliance on UD in this thesis reflects both an opportunity and a limitation. On the one hand, UD offers a linguistically principled, cross-linguistic framework that was readily available for our modeling. On the other hand, its design was not developed with topic inference in mind, but rather as a general-purpose annotation scheme. While UD has sometimes been critiqued for not adhering strictly to syntactic considerations (see Appendix A.2), we found this feature particularly advantageous for topic modeling: its treatment of entities, events, and semantic hierarchies between heads and dependents aligned well with the objective of modeling topic propagation. In light of our findings, one could even imagine developing an annotation scheme specifically tailored to support topic inference. Our results suggest that topics tend to propagate within nominal and clausal constructions but less so across constructions. Yet, even within multiword nominals or clauses, topic propagation was not guaranteed when modeled solely on individual word dependencies, which appears less intuitive. A possible way forward might be a hybrid annotation scheme that combines dependency and constituency principles, collapsing certain dependencies into higher-level spans such as multiword compound phrases, entity phrases, or event phrases. This would place even stronger emphasis on semantic phenomena at a more abstract, language-agnostic level. Alternatively, one could envision developing a semantic layer on top of UD. Exploring such extensions, however, remains an open avenue for future research, particularly given the substantial resources required to build and maintain frameworks such as UD.

5.3 Conclusion

This thesis has shown that syntactic structure, as encoded in Universal Dependencies, constitutes a valuable and under-exploited resource for topic modeling. By integrating syntactic information into the topic assignment process, the labeled UD-LDA introduces a transparent, linguistically motivated framework that improves interpretability without sacrificing model performance. The

labeled UD-LDA illustrates how syntactic structure can guide probabilistic topic inference, highlighting the interplay between form and meaning in text data. As both a methodological contribution and a conceptual bridge, this work provides a foundation for further research at the intersection of linguistics and applied text analysis – particularly in settings where data is scarce and interpretability is essential.

Appendices

Appendix A

Universal Dependencies framework

A.1 Overview of UD relation types

Table A.1: Descriptions and examples of UD relation types.

Label	Description	Example: head / <i>dependent</i>
acl	clausal modifier of noun	“Times Square got to be the place to <i>be</i> ”
acl:relcl	adnominal relative clause modifier	“as a lady who <i>went</i> alone”
advcl	adverbial clause modifier	“this hotel is very nice as you would <i>expect</i> ”
advmod	adverbial modifier	“ <i>judging</i> from their website it looked like a nice hotel”
amod	adjectival modifier	“the staff takes <i>good</i> care of their guests”
appos	appositional modifier	“small hotel , only 6 <i>rooms</i> per floor”
aux	auxiliary function word	“you <i>could</i> walk anywhere easily”
aux:pass	passive auxiliary	“the room <i>was</i> tastefully decorated ”
case	case marking element (e.g., preposition)	“it was good enough <i>for</i> me ”
cc	coordination	“easy to get to subway, theaters <i>and</i> museums ”
cc:preconj	preconjunct	“the food was <i>either</i> buffet or a sit down dinner”
ccomp	clausal complement	“no one explained how the elevator <i>worked</i> ”
compound	word-level compounding	“enough <i>counter</i> space for toiletries”
compound:prt	phrasal verb particle	“the breakfast sets you <i>up</i> for the day”

Table A.1 (continued): Descriptions and examples of UD relation types.

Label	Description	Example: head / <i>dependent</i>
conj	conjunct	“the room was clean and <i>comfortable</i> ”
cop	copula	“we <i>were</i> very pleased with everything”
csubj	clausal subject	“ <i>being</i> in a busy area is great ”
det	determiner	“such <i>a</i> great service ”
det:predet	predeterminer	“ <i>such</i> a great service ”
discourse	discourse element	“ <i>Oh</i> , ask for reservations”
expl	expletive (existential “there” or “it”)	“ <i>there</i> are lots of restaurants”
fixed	fixed grammatical expressions that behave like function words	“missed the show because of the mistake”
flat	flat multiword expression	“the Waldorf Astoria ”
goeswith	links parts of a word that are separated but should go together	“great experience <i>all</i> together ”
iobj	indirect object	“they gave <i>us</i> free drinks”
list	Used for lists of comparable items	“Madison Square Garden , Empire State <i>Building</i> , subway <i>stations</i> ”
mark	function word marking a clause as subordinate	“great location especially <i>if</i> you are interested in the Broadway”
nmod	nominal modifier	“the size of the <i>room</i> ”
nmod:npmo	noun phrase used as adverbial modifier	“the hotel <i>itself</i> was really nice”
nmod:poss	possessive nominal modifier	“ <i>our</i> room had a great view”
nmod:tmod	temporal nominal modifier	“two free bottles of water every <i>day</i> ”
nsubj	nominal subject	<i>they</i> gave us free drinks
nsubj:pass	passive nominal subject	“ <i>towels</i> were left on the floor”
nummod	numeric modifier	“ <i>two</i> free bottles of water every <i>day</i> ”
obj	direct object	“they gave us free <i>drinks</i> ”
obl	nominal functioning as a non-core (oblique) modifier of a predicate	“the subway got us to the <i>city</i> ”
obl:npmo	noun phrase as adverbial modifier	“my wife is six <i>months</i> pregnant ”
obl:tmod	temporal modifier	“we arrived late in the <i>morning</i> ”
parataxis	links constituents placed side by side	“they informed an engineer; no one ever <i>came</i> ”
root	root of the sentence	“our room had a great view”
vocative	marks dialogue participant	“ Thanks , <i>Marriott</i> !”
xcomp	open clausal complement	“the hotel was able to <i>upgrade</i> us”

A.2 Critique of the Universal Dependencies framework

As noted in the introductory chapter on the Universal Dependencies (UD) framework, the project has faced criticism, particularly from linguists. We acknowledge these critiques and aim to briefly reflect on them while justifying the relevance of the framework for our use case in topic modeling.

In the introduction, we outlined the six design criteria of UD (De Marneffe et al. 2021). These criteria serve as a useful lens to categorize the criticism, helping us identify which aspects are being challenged and assess their relevance to our specific application.

1. *“UD needs to be reasonably satisfactory on linguistic analysis grounds for individual languages — a journeyman’s universal grammar.*
2. *UD needs to be good for linguistic typology: It should bring out cross-linguistic parallelism across languages and language families.*
3. *UD must be suitable for rapid, consistent annotation by a human annotator.*
4. *UD must be easily comprehended and used by non-linguist users with prosaic needs.*
5. *UD must be suitable for computer parsing with high accuracy.*
6. *UD must support well downstream language understanding tasks, such as relation extraction, reading comprehension, machine translation, and so on.”*

A significant portion of the criticism directed at UD revolves around the **decision to make content words heads over function words**. This structural choice has sparked debate regarding its linguistic foundation and its implications for linguistic typology, particularly concerning criteria (1) and (2). Critics question whether this decision aligns with linguistic theory and whether the resulting dependencies effectively support cross-linguistic typological analyses.

Osborne & Gerdes (2019) argue that the binary classification of content and function words is not clear-cut. More importantly, they claim that this distinction is fundamentally semantic rather than syntactic, making it challenging to capture syntactic phenomena based on semantic

considerations. Similarly, Gerdes & Kahane (2016) criticize the alternation between semantic and syntactic arguments in UD annotation principles.

Osborne & Maxwell (2015) provide an overview of the historical treatment of function words in dependency grammars, examining whether the words UD classifies as function words have traditionally been considered subordinated or superordinated in linguistic grammars. Their focus is particularly on the relationships between prepositions and their nouns, auxiliary verbs and their content verbs, and determiners and their nouns.

They argue that there is no historical evidence for subordinating adpositions (prepositions and postpositions) to their nouns, since adpositions assign case to their nouns, not vice versa. This issue is illustrated with the sentence “we went to New York and had a great time”. In Universal Dependencies, *case(York, to)*, the preposition “to” is made dependent of “York” (see Figure 2.2). Syntactically, the preposition “to” licenses the presence of the noun compound “New York” and should therefore be considered the head rather than the dependent. Consequently, the annotation guideline to subordinate prepositions to the content words they introduce is in contrast to linguistic theory (Osborne & Maxwell 2015).

UD also subordinates auxiliary verbs, such as “has” in *aux(traveled, has)*, classifying them as function words. This contradicts the subcategorization principle, which is assumed to operate downward in the syntactic hierarchy. According to Osborne & Gerdes (2019), a word W1 subcategorizes W2 if W1 is more flexible in form than W2. For instance, while “had traveled” or “will have traveled” are valid constructions, “has travels” is not, indicating that the auxiliary verb “has” is syntactically more flexible and should therefore be superordinated to “traveled”. However, in UD, subordinating the auxiliary verb “has” leads to subcategorization operating upward in the dependency tree hierarchy, from the dependent “has” to the head “traveled”.

As an interim conclusion, Criterion 1 – requiring UD to be a reasonably satisfactory framework for linguistic analysis of individual languages – is subject to justified criticism. Specifically, the decision to make content words heads over function words is not standard practice in purely syntactic dependency frameworks. However, this issue is not insurmountable, as methods exist to convert UD into syntactically motivated annotations. Gerdes et al. (2018) propose Surface-Syntactic Universal Dependencies (SUD), an annotation scheme based solely on syntactic criteria,

and provide guidelines for converting UD into SUD. Thus, researchers interested in syntactic phenomena per se could opt for SUD instead of UD.

Osborne & Gerdes (2019) further argue that the UD's goal of promoting parallelism across languages (Criterion 2) conflicts with the objectives of typological studies, which focus on examining structural differences among typologically diverse languages. In this context, Dependency Distance – the number of intervening words between a head and its dependent – serves as a relevant metric for typologists to assess syntactic complexity across languages. However, the relatively flat structure of UD results in higher Dependency Distances compared to more layered, purely syntactic dependency annotations, potentially introducing bias in such analyses Osborne & Gerdes (2019), Yan & Liu (2022).

Other scholars, such as Futrell et al. (2020), argue that UD provide a sufficiently accurate representation of the cross-linguistic syntax to study dependency distance. The extent to which UD is suitable for analyzing linguistic phenomena such as dependency distance remains an open question.

Criteria 3, 4, and 5 have faced less criticism. Osborne & Gerdes (2019) do not provide strong arguments regarding whether human annotation consistency, comprehensibility for non-linguists, or parser accuracy are significantly better or worse compared to purely syntactic annotation formats.

Regarding Criterion 6, Osborne & Gerdes (2019) leave an open question. Although they critique UD from a linguistic perspective, they do not dispute its usefulness for downstream applications. Their criticism primarily addresses theoretical linguistic concerns rather than the practical applicability of the UD framework in computational tasks.

For our project, the key question is not whether UD provides a complete and linguistically precise representation of syntactic structures but rather whether the information it encodes is valuable for our specific application. In this regard, our approach is similar to that of Atalay et al. (2023), who use UD to study syntactic surprise in marketing communication. Their focus is not on evaluating the linguistic accuracy of UD, but on leveraging its structure to quantify the unexpectedness of syntax in text data. Likewise, our goal is not to scrutinize the exact syntactic constructions of user-generated content, but to enhance topic modeling by

incorporating additional structural information from UD. To do so effectively, we must first understand what kind of information UD encapsulates and what it does not.

A central characteristic of UD is its practice of making content words heads over function words, which results in content words being positioned higher and function words lower in the dependency tree hierarchy. Croft et al. (2017) support this approach, arguing that it is typologically appropriate to establish direct links between content words. These links are particularly relevant to topic modeling: if key thematic words in a sentence are directly connected, it becomes easier to trace how topics emerge and evolve in textual data.

To conclude, from a purely linguistic perspective, UD's blend of semantic and syntactic principles may be problematic. However, for our project, this characteristic is an advantage. By prioritizing content words and directly linking them within sentences, UD offers a structural representation that aligns well with the goals of topic modeling.

Appendix B

MCMC for the UD-LDA model

The following description of the MCMC is based on the derivation of Büschken & Allenby (2020) for the AT-LDA. Their approach is based on the sequential order of words in text and models topic carryover from one word to the next within the observed sequence. In the case of the UD-LDA, their representation is adapted in such a way that it allows for topic propagation between head-dependent word pairs within sentences. As stated above, the joint distribution of knowns and unknowns in the UD-LDA factorizes according to the formula presented in 3.5:

$$\begin{aligned}
 p(\{w\}_d, \{z\}_d, \{\zeta\}_d, \phi, \theta_d, \beta, x_d, \alpha, \tau, \bar{\beta}, \Sigma_\beta) \propto \\
 \prod_{s=1}^{S_d} \left[p(w_{s,root} | \phi, z_{s,root}) \times p(z_{s,root} | \theta_d) \times \right. \\
 \left. \prod_{\substack{n=1 \\ n \neq \text{root}}}^{N_s} \left[p(w_{s,n} | \phi, z_{s,n}, z_{s,h_n}, \zeta_{s,n}) \times p(z_{s,n} | z_{s,h_n}, \theta_d, \zeta_{s,n}) \times p(\zeta_{s,n} | x_{s,n}, \beta, z_{s,h_n}) \right] \right] \times \\
 p(\phi | \tau) \times p(\theta_d | \alpha) \times p(\beta | \bar{\beta}, \Sigma_\beta) \times \\
 p(\tau) \times p(\alpha) \times p(\bar{\beta}) \times p(\Sigma_\beta)
 \end{aligned}$$

This requires sampling steps for the following parameters:

- ϕ word distributions per topic
- θ topic distributions per document
- β topic propagation regression coefficients

z topic indicator on word level

ζ topic propagation indicator on word level

B.1 Draw of ϕ

The draw of ϕ_t is not changed compared to the standard LDA model (Blei et al. 2003) or the AT-LDA model (Büschken & Allenby 2020):

$$\begin{aligned} p(\phi_t | else) &\propto \prod_{d=1}^D \prod_{s=1}^{S_d} \prod_{n=1}^{N_s} p(w_n | \phi_t, z_n = t) \times p(\phi_t | \tau) \quad \forall t \\ &\sim \text{Dirichlet}(C_t^{WT} + \tau) \end{aligned} \quad (\text{B.1})$$

Where C_t^{WT} are the (given/observed) word topic counts. For the update of ϕ_t we treat z as given - meaning that the update of ϕ_t is conditionally independent of ζ . For C_t^{WT} we therefore do not distinguish whether observations of z are the result of topic propagation. The conjugate setup allows us to generate direct draws from the full conditional of $p(\phi_t | else)$.

B.2 Draw of θ

As outlined by Büschken & Allenby (2020), the update of θ_d is also a conjugate setup combining a multinomial likelihood and a Dirichlet prior. They also mention that for the update of θ_d only those topic assignments downstream of θ_d are relevant (i.e., only those topic assignments that are not the result of topic propagation):

$$\begin{aligned} p(\theta_d | else) &\propto \prod_{s=1}^{S_d} p(z_{s,root} | \theta_d) \prod_{\substack{n=1 \\ n \neq \text{root}}}^{N_s} p(z_{s,n} | z_{s,h_n}, \theta_d, \zeta_{s,n}) \times p(\theta_d | \alpha) \\ &= \prod_{n:\zeta=0}^{N_d} p(z_n | \theta_d) \times p(\theta_d | \alpha) \times \prod_{n:\zeta=1}^{N_d} 1 \\ &= \prod_{n:\zeta=0}^{N_d} p(z_n | \theta_d) \times p(\theta_d | \alpha) \\ &\sim \text{Dirichlet}(C_d^{TD} + \alpha) \end{aligned} \quad (\text{B.2})$$

Where C_d^{TD} is the topic-document matrix counting those topic assignments per document for

which the condition $\zeta_n = 0$ is true. The conjugate setup allows us to generate direct draws of the full conditional $p(\theta_d | \text{else})$.

B.3 Draw of β

By definition, the root word of a sentence is the only word in a sentence that does not have another word as its head. Hence, the root word's topic assignment cannot be the result of topic propagation, which means that $\zeta_{s, \text{root}} = 0$. Consequently, for the update of β the root word of each sentence is ignored:

$$\begin{aligned}
 p(\beta | \text{else}) &\propto \\
 &\prod_{d=1}^D \prod_{s=1}^{S_d} \prod_{\substack{n=1 \\ n \neq \text{root}}}^{N_s} p(\zeta_{d,s,n} | x_{d,s,n}, \beta, z_{d,s,h_n}) \times p(\beta | \bar{\beta}, \Sigma_\beta) = \\
 &\prod_{d=1}^D \prod_{s=1}^{S_d} \prod_{\substack{n=1 \\ n \neq \text{root}}}^{N_s} \frac{\exp[x'_{d,s,n} \beta]}{1 + \exp[x'_{d,s,n} \beta]} \times p(\beta | \bar{\beta}, \Sigma_\beta) \\
 &\sim \text{Binomial}(\psi_{d,s,n}) \times \text{MVN}(\bar{\beta}, \Sigma_\beta)
 \end{aligned} \tag{B.3}$$

For the update of β we have a non-conjugate setup consisting of a binomial likelihood and a weakly informative multivariate normal prior. The full conditional is not of a known form, which means that we cannot directly generate draws from this distribution. Therefore, the draw of β is accomplished by a Random Walk Metropolis Hastings step.

B.4 Draw of z_n and ζ_n

For the UD-LDA, it is possible to employ the joint “single-move” sampling step for the topic and propagation indicator, z_n and ζ_n , as outlined in Büschken & Allenby (2020) for the AT-LDA. The distinction between UD-LDA and AT-LDA is that we no longer model first-order topic propagation within the observed sequence, specifically from word w_{n-1} to word w_n . Instead, the UD-LDA models topic propagation along the branches of syntactic parse trees, i.e. from the head word w_{h_n} to the focal word w_n . Whereby the word w_n in turn can inherit the topic to its syntactic “children” $\{w_{c_n}\}$. This implies that the propagation indicator ζ_n is a function of

the topic assignment of the syntactic head word z_{h_n} . Similarly, the downstream propagation probabilities in the syntactic hierarchy, denoted by ζ_{c_n} , are a function of z_n . Thus, it follows that these hierarchical dependencies must be considered when drawing inference on z_n and ζ_n . In order to derive the joint posterior distribution of z_n and ζ_n , we can remove all elements from 3.5 that do not depend on z_n or ζ_n . We also treat $z_{h_n}, \zeta_{h_n, z_{c_n}}$ and ζ_{c_n} as given. The subscript h_n represents the (one) head word of the focal word, w_n . The subscript c_n represents all direct and indirect (syntactic) children of the focal word w_n whose (downstream) topic assignments are dependent on z_n or z_{h_n} .

$$\begin{aligned}
p(z_n = t, \zeta_n | z_{h_n}, \zeta_{c_n}, else) &\propto \\
p(w_n, w_{c_n} | \phi, z_n = t, z_{h_n}, \zeta_n, \zeta_{c_n}) &\times p(z_n = t | z_{h_n}, \theta_d, \zeta_n) \times \\
p(\zeta_n | z_{h_n}, \psi_n) &\times p(\zeta_{c_n} | z_n = t, \psi_{c_n})
\end{aligned} \tag{B.4}$$

For the joint distribution $p(z_n = t, \zeta_n | z_{h_n}, \zeta_{c_n}, else)$ we distinguish between two cases $\zeta_n = 1$ and $\zeta_n = 0$. Therefore, it can be seen that the topic assignment of the focal word can either be inherited from its head word or drawn from the document-specific prior θ_d . For the likelihood, there are also two cases with respect to a focal word's single syntactic "child" as indicated by $\zeta_{c_n} = 1$ and $\zeta_{c_n} = 0$. Depending on whether the topic assignment is further inherited, z_{c_n} and w_{c_n} must be taken into account. This results in 2×2 cases for the joint draw for the joint distribution of z_n, ζ_n :

$$p(z_n = t, \zeta_n = 0 | \zeta_{c_n} = 0, else) \propto \phi_t^{(w_n)} \times (1 - \psi_{z_{h_n}}^{(w_n)}) \times \theta_{d,t} \times (1 - \psi_t^{(w_{c_n})})$$

$$p(z_n = t, \zeta_n = 0 | \zeta_{c_n} = 1, else) \propto \phi_t^{(w_n)} \times (1 - \psi_{z_{h_n}}^{(w_n)}) \times \theta_{d,t} \times \phi_t^{(w_{c_n})} \times \psi_t^{(w_{c_n})}$$

$$p(z_n = z_{h_n}, \zeta_n = 1 | \zeta_{c_n} = 0, else) \propto \phi_{z_{h_n}}^{(w_n)} \times \psi_{z_{h_n}}^{(w_n)} \times (1 - \psi_{z_{h_n}}^{(w_{c_n})})$$

$$p(z_n = z_{h_n}, \zeta_n = 1 | \zeta_{c_n} = 1, else) \propto \phi_{z_{h_n}}^{(w_n)} \times \psi_{z_{h_n}}^{(w_n)} \times \phi_{z_{h_n}}^{(w_{c_n})} \times \psi_{z_{h_n}}^{(w_{c_n})}$$

In essence, the variable ζ_n serves to distinguish whether topic propagation is established ($\zeta_n = 1; z_n = z_{h_n}$) or not. In the absence of a topic propagation ($\zeta_n = 0$) the topic assignment z_n is generated from θ_d . Hence, the joint draw of ζ_n, z_n originates from $T + 1$ multinomial probabilities. A more detailed derivation of this joint move of ζ_n and z_n can be found in Büschken & Allenby (2020).

Appendix C

Measuring topic coherence

Coherence can be considered a measure of within-topic quality. Coherent topics suggest that the N most likely words of a topic should be perceived by humans as fitting together in a natural or reasonable way. Much of the work on coherence has focused on **quantifying this perception** within the context of topic modeling.

To do so, researchers have used human judgment tasks as a benchmark for measuring perceived coherence. Chang et al. (2009) addressed this issue through a word intrusion task, a variant of the 'odd-one-out' game, where participants identify the randomly inserted word from a list of the N most likely words generated by the topic model. If a topic is highly coherent – meaning its words naturally fit together – detecting the intruder word should be relatively easy. Their findings demonstrated that even non-experts in topic modeling have an intuitive sense of whether a word list appears semantically coherent. Newman et al. (2010) extended this idea by introducing the concept of topic “*usefulness*” in downstream tasks. In their study, participants rated the usefulness of topic word lists on a scale from 1 to 3. While they did not provide a strict definition of “*useful*”, they instructed participants “*it is some combination of coherent, meaningful, interpretable, words-are-related, subject-heading-like, something-you-could-easily label*” (Newman et al. 2010). Mimno et al. (2011) further refined human evaluation by instructing experts to rate topics as 'good' if the words could be grouped into a single concept. Similarly, Aletras & Stevenson (2013) defined a topic as coherent if all its terms were associated with a common theme, advising raters to assess whether topics were “*semantically coherent*,

meaningful, and interpretable". While authors define topic coherence in slightly different ways, they share a common idea: a topic is coherent if its N most likely words can be subsumed under a common theme.

Several scholars have proposed quantitative coherence measures that automatically assess the semantic quality of topics and have validated these metrics against human judgments. Röder et al. (2015) offer a conceptual framework that outlines how such coherence measures are constructed (see Table C.1). According to their framework, coherence measures typically consist of four key components: (1) segmentation, (2) confirmation measure, (3) word probabilities or word vectors and (4) aggregation.

In the first step, **segmentation**, word pairs are generated from the list of the top N words in a topic. The most commonly used segmentation scheme is S_{one}^{one} , which forms all possible word pairs. For N top words, this results in $(N \cdot (N - 1))/2$ word pairs. In the second step, a **confirmation measure** is computed for each word pair. This measure is based either on (conditional) word probabilities or on the similarity of word vectors. Measures based on word probabilities typically rely on Boolean co-occurrence, where the presence or absence of a word is recorded within a specified scope – such as a document, sliding window, or context window. These methods count how often topic words appear together within that scope. In contrast, word vectors attempt to encode contextual usage patterns beyond simple co-occurrence counts, offering a more nuanced representation of word meaning. In the final step, **aggregation**, the individual confirmation scores for all word pairs are combined into a single coherence score for topic t , typically by computing their arithmetic mean. Model-level coherence scores are then obtained by averaging across topics. Since the compared coherence measures all use the same aggregation method, this component is not included in Table C.1.

Table C.1 summarizes the most commonly used coherence measures according to the criteria outlined above. Each of these metrics has been evaluated in prior studies through comparisons with human judgments, providing empirical support for their validity.

Generally, the five depicted types of coherence measures can be distinguished based on two dimensions: (1) they either utilize **direct or indirect** confirmation measures, (2) they are based on an **intrinsic or extrinsic** reference corpus (see Table C.1, second column).

The first three metrics – LCP, PMI, and NPMI – use direct confirmation measures. These metrics evaluate coherence by **directly** comparing pairwise word counts, based on co-occurrence within a defined context such as a document or a sliding window. A high coherence score is achieved when two words frequently appear together in the same context, thereby increasing the co-occurrence numerator.

In contrast, the last two metrics – Distributional Similarity and Word2Vec Similarity – rely on indirect confirmation measures. Instead of word counts, they compare word vectors that encode contextual usage, for example, within a window of ± 5 words. A high coherence score is obtained when two words have similar vectors, indicating that they tend to appear in comparable local contexts. In this way, these metrics assess semantic relatedness **indirectly**, by analyzing the similarity of the environments in which the words occur.

Coherence Measure Author	Type	Segmentation (S)	Confirmation Measure (M)	Word Probabilities (P) or Word Vectors
LCP (UMASS) Mimno et al. (2011)	direct intrinsic	S_{pre}^{one}	$m_{lc}(S_i) = \log \frac{P(W', W^*) + \epsilon}{P(W^*)}$	$P_{bd} \triangleq$ boolean document
PMI (UCI) Newman et al. (2010)	direct extrinsic	S_{one}^{one}	$m_{lr}(S_i) = \log \frac{P(W', W^*) + \epsilon}{P(W')P(W^*)}$	$P_{sw(10)} \triangleq$ boolean sliding window of size 10
NPMI Bouma (2009)	direct extrinsic	S_{one}^{one}	$m_{nlr}(S_i) = \log \frac{m_{lr}(S_i)}{-\log(P(W', W^*) + \epsilon)}$	$P_{sw(10)} \triangleq$ boolean sliding window of size 10
Distributional Similarity Aletras & Stevenson (2013)	indirect extrinsic	S_{one}^{one}	$\tilde{m}_{sim(nlr, \gamma)}(W', W^*) = s_{cos}(\vec{v}_{nlr, \gamma}(W'), \vec{u}_{nlr, \gamma}(W^*))$	$P_{cw(5)} \triangleq$ boolean context window of size ± 5
Word2Vec Similarity O'callaghan et al. (2015)	indirect extrinsic	S_{one}^{one}	$\tilde{m}_{sim(w2v)}(W', W^*) = s_{cos}(\vec{v}_{w2v}(W'), \vec{u}_{w2v}(W^*))$	latent vector representations (dim=200) learned via word2vec model and Skip-gram context window of size ± 5

Table C.1: Description of commonly used coherence measures. Adapted from Röder et al. (2015).

Bibliography

- Abney, S. (1997), Part-of-speech tagging and partial parsing, *in* ‘Corpus-based methods in language and speech processing’, Springer, pp. 118–136.
- Aletras, N. & Stevenson, M. (2013), Evaluating topic coherence using distributional semantics, *in* ‘Proceedings of the 10th international conference on computational semantics (IWCS 2013)–Long Papers’, pp. 13–22.
- Atalay, A. S., Kihal, S. E. & Ellsaesser, F. (2023), ‘Creating effective marketing messages through moderately surprising syntax’, *Journal of Marketing* **87**(5), 755–775.
- Balikas, G., Amoualian, H., Clausel, M., Gaussier, E. & Amini, M. R. (2016), Modeling topic dependencies in semantically coherent text spans with copulas, *in* ‘Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers’, The COLING 2016 Organizing Committee, pp. 1767–1776.
URL: <https://aclanthology.org/C16-1166>
- Berger, J., Humphreys, A., Ludwig, S., Moe, W. W., Netzer, O. & Schweidel, D. A. (2020), ‘Uniting the tribes: Using text for marketing insight’, *Journal of marketing* **84**(1), 1–25.
- Bird, S., Klein, E. & Loper, E. (2009), *Natural language processing with Python: analyzing text with the natural language toolkit*, " O'Reilly Media, Inc."
- Blei, D. M. (2012), ‘Probabilistic topic models’, *Communications of the ACM* **55**(4), 77–84.
- Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003), ‘Latent dirichlet allocation’, *the Journal of machine Learning research* **3**, 993–1022.

- Bouma, G. (2009), ‘Normalized (pointwise) mutual information in collocation extraction’, *Proceedings of GSCL* **30**, 31–40.
- Boyd-Graber, J. & Blei, D. (2008), ‘Syntactic topic models’, *Advances in neural information processing systems* **21**.
- Brown, T. B. (2020), ‘Language models are few-shot learners’, *arXiv preprint arXiv:2005.14165* .
- Büschken, J. & Allenby, G. M. (2016), ‘Sentence-based text analysis for customer reviews’, *Marketing Science* **35**(6), 953–975.
- Büschken, J. & Allenby, G. M. (2020), ‘Improving text analysis using sentence conjunctions and punctuation’, *Marketing Science* **39**(4), 727–742.
- Büschken, J. & Allenby, G. M. (2025), ‘A topic change point model for phrase-based topic inference r. 3’, *International Journal of Research in Marketing* .
- Casella, G. & George, E. I. (1992), ‘Explaining the gibbs sampler’, *The American Statistician* **46**(3), 167–174.
- Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. & Blei, D. (2009), ‘Reading tea leaves: How humans interpret topic models’, *Advances in neural information processing systems* **22**, 1–9.
- Croft, W., Nordquist, D., Looney, K. & Regan, M. (2017), Linguistic typology meets universal dependencies., *in* ‘TLT’, pp. 63–75.
- Darling, W. M. (2011), A theoretical and practical implementation tutorial on topic modeling and gibbs sampling, *in* ‘Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies’, pp. 642–647.
- Darling, W. M., Paul, M. & Song, F. (2012), Unsupervised part-of-speech tagging in noisy and esoteric domains with a syntactic-semantic bayesian hmm, *in* ‘Proceedings of the Workshop on Semantic Analysis in Social Media’, pp. 1–9.
- Darling, W. M. & Song, F. (2013), ‘Probabilistic topic and syntax modeling with part-of-speech lda’, *arXiv preprint arXiv:1303.2826* .

- De Marneffe, M.-C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J. & Manning, C. D. (2014), Universal stanford dependencies: A cross-linguistic typology., *in* 'LREC', Vol. 14, pp. 4585–4592.
- De Marneffe, M.-C., MacCartney, B., Manning, C. D. et al. (2006), Generating typed dependency parses from phrase structure parses., *in* 'Lrec', Vol. 6, pp. 449–454.
- De Marneffe, M.-C. & Manning, C. D. (2008), The stanford typed dependencies representation, *in* 'Coling 2008: proceedings of the workshop on cross-framework and cross-domain parser evaluation', pp. 1–8.
- De Marneffe, M.-C., Manning, C. D., Nivre, J. & Zeman, D. (2021), 'Universal dependencies', *Computational linguistics* **47**(2), 255–308.
- Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. & Harshman, R. (1990), 'Indexing by latent semantic analysis', *Journal of the American society for information science* **41**(6), 391–407.
- Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2019), Bert: Pre-training of deep bidirectional transformers for language understanding, *in* 'Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)', pp. 4171–4186.
- Fan, A., Doshi-Velez, F. & Miratrix, L. (2019), 'Assessing topic model relevance: Evaluation and informative priors', *Statistical Analysis and Data Mining: The ASA Data Science Journal* **12**(3), 210–222.
- Futrell, R., Levy, R. P. & Gibson, E. (2020), 'Dependency locality as an explanatory principle for word order', *Language* **96**(2), 371–412.
- Futrell, R., Mahowald, K. & Gibson, E. (2015), 'Large-scale evidence of dependency length minimization in 37 languages', *Proceedings of the National Academy of Sciences* **112**(33), 10336–10341.

- Gerdes, K., Guillaume, B., Kahane, S. & Perrier, G. (2018), Sud or surface-syntactic universal dependencies: An annotation scheme near-isomorphic to ud, in ‘Universal dependencies workshop 2018’.
- Gerdes, K. & Kahane, S. (2016), Dependency annotation choices: Assessing theoretical and practical issues of universal dependencies, in ‘LAW X (2016) The 10th Linguistic Annotation Workshop: 131’.
- Gilks, W. R., Richardson, S. & Spiegelhalter, D. (1995), *Markov chain Monte Carlo in practice*, CRC press.
- Griffiths, T. L. & Steyvers, M. (2004), ‘Finding scientific topics’, *Proceedings of the National academy of Sciences* **101**(suppl_1), 5228–5235.
- Griffiths, T., Steyvers, M., Blei, D. & Tenenbaum, J. (2004), ‘Integrating topics and syntax’, *Advances in neural information processing systems* **17**.
- Gruber, A., Weiss, Y. & Rosen-Zvi, M. (2007), Hidden topic markov models, in ‘Artificial intelligence and statistics’, PMLR, pp. 163–170.
- Heinrich, G. (2008), Parameter estimation for text analysis, Technical report, Citeseer.
- Hofmann, T. (1999), Probabilistic latent semantic indexing, in ‘Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval’, pp. 50–57.
- Hong, J. & Hoban, P. R. (2022), ‘Writing more compelling creative appeals: A deep learning-based approach’, *Marketing Science* **41**(5), 941–965.
- Hruschka, H. (2014), ‘Linking multi-category purchases to latent activities of shoppers: analysing market baskets by topic models’, *Marketing: ZFP–Journal of Research and Management* **36**(4), 267–273.
- Jan Wijffels, Milan Straka, J. S. (2022), *udpipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the ‘UDPipe’ ‘NLP’ Toolkit*. R package version 0.8.11.
URL: <https://bnosac.github.io/udpipe/en/index.html>

- Jiang, J. (2009), Modeling syntactic structures of topics with a nested hmm-lda, in ‘2009 Ninth IEEE International Conference on Data Mining’, IEEE, pp. 824–829.
- Johansson, R. & Nugues, P. (2007), Extended constituent-to-dependency conversion for english, in ‘16th Nordic Conference of Computational Linguistics’, University of Tartu, pp. 105–112.
- Jurafsky, D. & Martin, J. H. (2021), ‘Speech and language processing (3rd edition draft)’.
URL: https://web.stanford.edu/jurafsky/slp3/old_dec21/indexdec21.html
- Kronrod, A. et al. (2022), ‘Language research in marketing’, *Foundations and Trends® in Marketing* **16**(3), 308–421.
- Labrecque, L. I., Swani, K. & Stephen, A. T. (2020), ‘The impact of pronoun choices on consumer engagement actions: Exploring top global brands’ social media communications’, *Psychology & Marketing* **37**(6), 796–814.
- Lee, T. Y. & Bradlow, E. T. (2011), ‘Automated marketing research using online customer reviews’, *Journal of Marketing Research* **48**(5), 881–894.
- Li, J., Zhang, M., Wang, J., Zhang, M. & Zhang, Y. (2025), ‘Universal constituency treebanking and parsing: A pilot study’, *Computer Speech & Language* p. 101826.
- Ludwig, S., De Ruyter, K., Friedman, M., Brüggem, E. C., Wetzels, M. & Pfann, G. (2013), ‘More than words: The influence of affective content and linguistic style matches in online reviews on conversion rates’, *Journal of marketing* **77**(1), 87–103.
- Marcus, M., Santorini, B. & Marcinkiewicz, M. A. (1993), ‘Building a large annotated corpus of english: The penn treebank’, *Computational linguistics* **19**(2), 313–330.
- Martin, F. & Johnson, M. (2015), More efficient topic modelling through a noun only approach, in ‘Proceedings of the Australasian Language Technology Association Workshop 2015’, pp. 111–115.
- McDonald, R., Nivre, J., Quirnbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., Hall, K., Petrov, S., Zhang, H., Täckström, O. et al. (2013), Universal dependency annotation

- for multilingual parsing, in ‘Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)’, pp. 92–97.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. (2013), ‘Distributed representations of words and phrases and their compositionality’, *Advances in neural information processing systems* **26**.
- Miller, J. & Miller, J. (2011), *A Critical Introduction to Syntax*, Bloomsbury Critical Introductions to Linguistics, Bloomsbury Academic.
- URL:** <https://books.google.de/books?id=a0UQHxUzR9AC>
- Mimno, D., Wallach, H., Talley, E., Leenders, M. & McCallum, A. (2011), Optimizing semantic coherence in topic models, in ‘Proceedings of the 2011 conference on empirical methods in natural language processing’, pp. 262–272.
- Netzer, O., Feldman, R., Goldenberg, J. & Fresko, M. (2012), ‘Mine your own business: Market-structure surveillance through text mining’, *Marketing Science* **31**(3), 521–543.
- Newman, D., Noh, Y., Talley, E., Karimi, S. & Baldwin, T. (2010), Evaluating topic models for digital libraries, in ‘Proceedings of the 10th annual joint conference on Digital libraries’, pp. 215–224.
- Nivre, J., De Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N. et al. (2016), Universal dependencies v1: A multilingual treebank collection, in ‘Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)’, pp. 1659–1666.
- Nivre, J., De Marneffe, M.-C., Ginter, F., Hajič, J., Manning, C. D., Pyysalo, S., Schuster, S., Tyers, F. & Zeman, D. (2020), ‘Universal dependencies v2: An evergrowing multilingual treebank collection’, *arXiv preprint arXiv:2004.10643*.
- Osborne, T. & Gerdes, K. (2019), ‘The status of function words in dependency grammar: A critique of universal dependencies (ud)’, *Glossa: a journal of general linguistics (2016-2021)*.

- Osborne, T. & Maxwell, D. (2015), A historical overview of the status of function words in dependency grammar, *in* ‘Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)’, pp. 241–250.
- O’callaghan, D., Greene, D., Carthy, J. & Cunningham, P. (2015), ‘An analysis of the coherence of descriptors in topic modeling’, *Expert Systems with Applications* **42**(13), 5645–5657.
- Pancer, E., Chandler, V., Poole, M. & Noseworthy, T. J. (2019), ‘How readability shapes social media engagement’, *Journal of consumer psychology* **29**(2), 262–270.
- Petrov, S., Das, D. & McDonald, R. (2011), ‘A universal part-of-speech tagset’, *arXiv preprint arXiv:1104.2086* .
- Qi, P., Zhang, Y., Zhang, Y., Bolton, J. & 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’.
- URL:** <https://nlp.stanford.edu/pubs/qi2020stanza.pdf>
- Reisenbichler, M. & Reutterer, T. (2019), ‘Topic modeling in marketing: recent advances and research opportunities’, *Journal of Business Economics* **89**(3), 327–356.
- Rocklage, M. D., He, S., Rucker, D. D. & Nordgren, L. F. (2023), ‘Beyond sentiment: The value and measurement of consumer certainty in language’, *Journal of Marketing Research* **60**(5), 870–888.
- Röder, M., Both, A. & Hinneburg, A. (2015), Exploring the space of topic coherence measures, *in* ‘Proceedings of the eighth ACM international conference on Web search and data mining’, pp. 399–408.
- Schweidel, D. A. & Moe, W. W. (2014), ‘Listening in on social media: A joint model of sentiment and venue format choice’, *Journal of marketing research* **51**(4), 387–402.
- Silveira, N., Dozat, T., de Marneffe, M.-C., Bowman, S., Connor, M., Bauer, J. & Manning, C. D. (2014), A gold standard dependency corpus for English, *in* ‘Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)’.

- Steyvers, M. & Griffiths, T. (2007), 'Probabilistic topic models', *Handbook of latent semantic analysis* **427**(7), 424–440.
- Straka, M., Hajic, J. & Straková, J. (2016), Udpipes: trainable pipeline for processing conll-u files performing tokenization, morphological analysis, pos tagging and parsing, in 'Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)', pp. 4290–4297.
- Straka, M. & Straková, J. (2017), Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe, in 'Proceedings of the CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies', pp. 88–99.
- Tesnière, L. (1959), 'Eléments de syntaxe structurale', Paris, Klincksieck **5**.
- Timoshenko, A. & Hauser, J. R. (2019), 'Identifying customer needs from user-generated content', *Marketing Science* **38**(1), 1–20.
- Tirunillai, S. & Tellis, G. J. (2014), 'Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent dirichlet allocation', *Journal of marketing research* **51**(4), 463–479.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. (2017), 'Attention is all you need', *Advances in neural information processing systems* **30**.
- Wallach, H. M. (2006), Topic modeling: beyond bag-of-words, in 'Proceedings of the 23rd international conference on Machine learning', pp. 977–984.
- Wallach, H. M. (2008), Structured topic models for language, PhD thesis, University of Cambridge Cambridge, UK.
- Wang, C., Blei, D. M. & Fei-Fei, L. (2009), Simultaneous image classification and annotation., in 'CVPR', Vol. 1, p. 6.
- Wang, X., McCallum, A. & Wei, X. (2007), Topical n-grams: Phrase and topic discovery, with

- an application to information retrieval, in ‘Seventh IEEE international conference on data mining (ICDM 2007)’, IEEE, pp. 697–702.
- Yan, J. & Liu, H. (2022), ‘Semantic roles or syntactic functions: The effects of annotation scheme on the results of dependency measures’, *Studia Linguistica* **76**(2), 406–428.
- Yee Whye Teh, Michael I Jordan, M. J. B. & Blei, D. M. (2006), ‘Hierarchical dirichlet processes’, *Journal of the American Statistical Association* **101**(476), 1566–1581.
- Zeman, D. (2008), Reusable tagset conversion using tagset drivers., in ‘LREC’, Vol. 2008, pp. 28–30.
- Zeman, D., Hajic, J., Popel, M., Potthast, M., Straka, M., Ginter, F., Nivre, J. & Petrov, S. (2018), Conll 2018 shared task: Multilingual parsing from raw text to universal dependencies, in ‘Proceedings of the CoNLL 2018 Shared Task: Multilingual parsing from raw text to universal dependencies’, pp. 1–21.
- Zeman, D., Popel, M., Straka, M., Hajič, J., Nivre, J., Ginter, F., Luotolahti, J., Pyysalo, S., Petrov, S., Potthast, M., Tyers, F., Badmaeva, E., Gokirmak, M., Nedoluzhko, A., Cinková, S., Hajič jr., J., Hlaváčová, J., Kettnerová, V., Urešová, Z., Kanerva, J., Ojala, S., Missilä, A., Manning, C. D., Schuster, S., Reddy, S., Taji, D., Habash, N., Leung, H., de Marneffe, M.-C., Sanguinetti, M., Simi, M., Kanayama, H., de Paiva, V., Drozanova, K., Martínez Alonso, H., Çöltekin, Ç., Sulubacak, U., Uszkoreit, H., Macketanz, V., Burchardt, A., Harris, K., Marheinecke, K., Rehm, G., Kayadelen, T., Attia, M., Elkahky, A., Yu, Z., Pitler, E., Lertpradit, S., Mandl, M., Kirchner, J., Alcalde, H. F., Strnadová, J., Banerjee, E., Manurung, R., Stella, A., Shimada, A., Kwak, S., Mendonça, G., Lando, T., Nitisaroj, R. & Li, J. (2017), CoNLL 2017 shared task: Multilingual parsing from raw text to Universal Dependencies, in J. Hajič & D. Zeman, eds, ‘Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies’, Association for Computational Linguistics, Vancouver, Canada, pp. 1–19.
- URL:** <https://aclanthology.org/K17-3001/>

Erklärung an Eides statt

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht. Die Arbeit wurde bisher in gleicher oder ähnlicher Form keiner anderen Prüfungsbehörde vorgelegt und auch nicht veröffentlicht.

Die elektronische Ausfertigung der Arbeit habe ich bereits beim Prüfer eingereicht.

Ingolstadt, September 2025

Thomas Michael Rußer-Strobel