States, events, and generics: computational modeling of situation entity types

Dissertation
zur Erlangung des akademischen Grades
eines Doktors der Philosophie
der Philosophischen Fakultät
der Universität des Saarlandes

vorgelegt von
Annemarie Silke Friedrich
aus Herrenberg

Saarbrücken, 2017

Dekan der Philosophischen Fakultät II: Prof. Dr. Erich Steiner

Erstberichterstatter: Prof. Dr. Manfred Pinkal

Zweitberichterstatter: Prof. Dr. Alexis Palmer

Drittberichterstatter: Prof. Dr. Josef van Genabith

Tag der letzten Prüfungsleistung: 24. Februar 2017

Abstract

This dissertation addresses the computational modeling of *situation entity types* (Smith, 2003), an inventory of clause types capturing aspectual and semantic distinctions that are relevant for various natural language processing tasks including temporal discourse processing and information extraction. The focus of our work is on automatically identifying the situation entity types State, ("John is tall"), Event ("John won the race"), Generalizing Sentence ("John cycles to work") and Generic Sentence ("Elephants are mammals").

We create a large corpus of texts from a variety of genres and domains, annotating each clause with its situation entity type and with linguistic phenomena that we identify as relevant for distinguishing the types. Specifically, we mark each clause with its *lexical aspectual class*, which takes the values ***stative*** ("be," "know") or ***dynamic*** ("run," "win"), and whether the clause is ***episodic*** or ***habitual***, i.e., whether it refers to a particular event or whether it generalizes over situations. In addition, we annotate whether a clause's subject is ***generic*** or not, i.e., whether it refers to a kind ("dogs") or to a particular individual ("my dog"). Our human annotators achieve substantial agreement for all of these annotation tasks. Based on this corpus, we conduct a detailed corpus-linguistic study of situation entity type distributions and variation in inter-annotator agreement depending on the genre.

In the second part of this dissertation, we create computational models for each of the above mentioned classification tasks in a supervised setting, advancing the state-of-the-art in each case. We find a range of syntactic-semantic features including distributional information and corpus-based linguistic indicators to be helpful. Using a sequence labeling method, we are able to leverage discourse information in order to improve the recognition of genericity, which often cannot be decided without taking the sentences in the context into account. We show our models to perform robustly across domains. Our publicly available data set and implementation form the basis for future research on situation entity types and related aspectual phenomena, among others as a preprocessing step into various natural language processing tasks.

Kurzzusammenfassung

Die vorliegende Dissertation befasst sich mit der computergestützten Modellierung von *Situationstypen* (Smith, 2003), einem Inventar von Satz- bzw. Teilsatztypen, das aspektuelle und semantische Unterscheidungen erfasst, die für verschiedene Bereiche der maschinellen Sprachverarbeitung relevant sind. Zu diesen Bereichen zählen beispielsweise die Erkennung von temporalen Diskursrelationen und die Informationsextraktion. Der Fokus dieser Arbeit liegt auf der automatischen Identifikation der Situationstypen *Zustand*, (State, “John ist groß”), *Ereignis* (Event, “John gewann das Rennen”), *generalisierender Satz* (Generalizing Sentence, “John fährt mit dem Fahrrad zur Arbeit”) und *generischer Satz* (Generic Sentence, “Elefanten sind Säugetiere”).

Als Grundlage für die Untersuchung wurde ein Korpus von Texten mehrerer Genres und Domänen auf Teilsatzebene manuell mit Situationstypen und weiteren für die Unterscheidung dieser Typen relevanten linguistischen Phänomenen annotiert. Jeder Teilsatz wird mit den *Aktionsarten statisch* (“sein”, “wissen”) oder *dynamisch* (“laufen”, “gewinnen”) annotiert und als *episodisch* oder *habituell* eingestuft, d.h., ob er ein ein bestimmtes Ereignis oder eine Generalisierung über Situationen beschreibt. Außerdem wird für jedes Subjekt annotiert, ob es *generisch* ist, d.h., ob es auf eine natürliche Gattung (“Hunde”) oder auf ein bestimmtes Individuum (“mein Hund”) referiert. Bei allen manuellen Annotationsaufgaben wird eine substanzielle Übereinstimmung erreicht. Eine auf diesem Korpus basierende detaillierte korpuslinguistische Studie zeigt genreabhängige Variationen in der Verteilung der Situationstypen und bei der jeweils zwischen den Annotatoren erreichten Übereinstimmung.

Der zweite Teil der Arbeit beschreibt die computergestützte Modellierung der oben erwähnten Klassifikationsaufgaben mit Hilfe von überwachten Lernalgorithmen. Bei allen vier Klassifikationsaufgaben verbessern die hier vorgestellten Modelle den jeweiligen Stand der Technik. Dabei zeigt sich, dass eine Auswahl von syntaktisch-semantischen Attributen, unter anderem distributionelle Information und korpusbasierte linguistische Indikatoren, für die Modellierung geeignet ist. Außerdem wird eine Methode zur Annotation von Sequenzen eingesetzt, die Diskursinformation nutzt, um die Genauigkeit bei der Erkennung von generischen Ausdrücken zu verbessern. Dies ist oft – auch manuell – nur unter Einbeziehung weiterer Sätze aus dem lokalen Diskurskontext möglich. Die hier vorgestellten Modelle zeigen auch über Genre-Grenzen hinweg eine robuste Performanz. Ein frei verfügbares Datenset und die frei verfügbare Implementierung bieten Ansatzpunkte für weitere Forschung im Bereich von Situationstypen und von verwandten aspektuellen Phänomenen, wie zum Beispiel die Integration in Vorverarbeitungsschritte diverser maschineller Sprachverarbeitungssysteme.

Ausführliche Zusammenfassung

Wenn wir natürliche Sprache benutzen, um Informationen zu vermitteln, lokalisieren wir die Situationen, über die wir sprechen, in der Zeit, d.h. in Vergangenheit, Gegenwart oder Zukunft. Durch die Wahl einer geeigneten Zeitform (*Tempus*) können in vielen Sprachen der Zeitpunkt oder temporale Relationen ausgedrückt werden. Mithilfe des *Aspekts* besteht die Möglichkeit, Situationen aus verschiedenen temporalen Perspektiven darzustellen (Comrie, 1976; Smith, 1997). Die Linguistik, insbesondere die theoretische Semantik, unterscheidet in diesem Zusammenhang unter anderem zwischen *Zuständen*, *Ereignissen* und *Prozessen* (Vendler, 1957; Bach, 1986).

Die vorliegende Dissertation beschäftigt sich mit der computergestützen Verarbeitung der aspektuellen Merkmale der Teilsätze eines Textes. Alle Beispiele in (1) sind sprachliche Realisierungen, die sich auf dasselbe Ereignis beziehen, jedoch verschiedene Phasen dieses Ereignisses fokussieren. Der erste Satz (a) beschreibt das gesamte Ereignis inklusive Anfang und Ende, während (b) und (c) nur die mittlere Phase des Ereignisses, in der das Schiff in Bewegung ist, für den Leser "sichtbar" machen. Im Fall von (c) wird das eigentliche Ereignis sogar sprachlich als Zustand realisiert.

(1) (a) Das Schiff hatte sich bewegt. (gesamtes, abgeschlossenes Ereignis)
(b) Das Schiff bewegte sich. (Prozess; Ereignis, das gerade im Gang ist)
(c) Das Schiff war in Bewegung. (Zustand)

Wir können außerdem ausdrücken, ob eine Situation ein einmalig stattfindendes Ereignis darstellt (2a) oder ob wir über eine Regelmäßigkeit berichten (2b). So genannte *generische Sätze* können sowohl Situationen (2b) als auch Aussagen über die Instanzen einer Gattung (2c) verallgemeinern (Krifka et al., 1995).

(2) (a) Gestern ist er Fahrrad gefahren.
(b) Er fährt (für gewöhnlich) mit dem Fahrrad zur Arbeit.
(c) Studenten mögen Kaffee.

Die vorliegende Arbeit zeigt Methoden für die automatische Identifikation der von den Teilsätzen eines Textes ausgedrückten Situationstypen. Wie bereits mithilfe von Beispiel (1) beschrieben, hat diese aspektuelle Klassifikation zum Ziel, die *linguistische Darstellung* von Situationen zu erfassen, nicht deren "tatsächliche" Eigenschaften (Filip, 2012, Abschnitt 2.5.6). Die linguistische Darstellung von Ereignissen wird in dem hier vorgestellten Ansatz auf Teilsatzebene klassifiziert und unterscheidet zwischen den Fällen in (1a-c). Diese Aufgabenstellung ist orthogonal zu Arbeiten im Bereich des *semantischen Parsing*, bei welchem die Identifkation von Prädikat-Arguments-Strukturen von zentralem Interesse ist. Letztere repräsentieren die Bedeutung eines Satzes, indem sie die Konstituenten eines Satzes den jeweiligen thematischen Rollen zuordnen, d.h. beispielsweise dem Agens (Subjekt) oder dem Thema (Objekt). Semantisches Parsing und aspektuelle Klassifikation sind sich insofern ähnlich, dass beide die Repräsentation von Bedeutung mit dem Ziel des automatischen Textverstehens adressieren. In der Computerlinguistik existieren im Bereich des semantischen Parsing zahlreiche Arbeiten, die unterschiedliche Formalismen verwenden. Zu den meistverwendeten Formalismen zählen PropBank (Palmer et al., 2005), FrameNet (Ruppenhofer et al., 2006) und Abstract Meaning Representations (Banarescu et al., 2013). Im Bereich der aspektuellen Klassifikation gibt es im Gegensatz dazu bisher nur relativ wenige empirische Arbeiten. Für automatisches Textverstehen, inklusive der korrekten Interpretation von temporalen Relationen und der Unterscheidung von spezieller und generischer Information, sind beide Bereiche notwendig. Die jeweils von den beiden Ansätzen erfassten Informationen ergänzen sich dabei. Wenn wir beispielsweise den Satz "Sie winkte zum Abschied" direkt nach einem der drei Sätze in (1) lesen, gelangen wir zu unterschiedlichen Schlussfolgerungen, je nachdem, welcher Kontext gewählt wurde. Im Fall von (b) und (c) folgern wir, dass das Winken stattgefunden hat, während das Schiff in Bewegung war, während wir im Fall von (a) zu der – zugegebenermaßen pragmatisch etwas ungewöhnlichen – Interpretation gelangen, dass sich das Schiff in Bewegung setzte, wieder zum Stillstand kam und dann erst das Ereignis des Winkens stattfand. Dieses Beispiel zeigt, wie aspektuelle Formen temporale Relationen zwischen Situationen bestimmen. In Sprachen wie Mandarin, die kein Tempus verwenden, ist dies ein wichtiger Mechanismus, um temporale Strukturen zu interpretieren (Smith and Erbaugh, 2005).

Die vorliegende Arbeit behandelt außerdem die automatische Erkennung von zwei Arten von generischen Ausdrücken. Bei dem ersten dieser beiden linguistischen Phänomene handelt es sich um *habituelle Aussagen*, d.h. um Sätze wie (3b), die Situationen betreffende Regelmäßigkeiten ausdrücken. Im Gegensatz dazu berichtet (3a) von einem bestimmten einmaligen Ereignis. Habituelle Sätze erlauben auch Ausnahmen: Satz (3b) wird auch dann noch als wahr eingestuft, wenn das Schiff nur in den meisten Jahren frisch gestrichen wird.

(3) (a) Sie strichen das Schiff an. (einmaliges Ereignis)
(b) Sie streichen das Schiff einmal im Jahr. (Regularität: habituell)

Bei dem zweiten linguistischen Phänomen handelt es sich um die *Referenz auf eine natürliche Gattung* wie in (4b). Diese Art von generischen Sätzen ist abzugrenzen von Sätzen wie (4a), die Aussagen über ein bestimmtes Objekt treffen.

(4) (a) Die RMS Titanic war ein britisches Passagierschiff. (Referenz auf Objekt)
(b) Schiffe werden im Allgemeinen auf Grund ihrer Größe, Form und Ladungs- und
Passagierkapazität von Booten unterschieden. (Referenz auf Art / Klasse)

Generische Sätze wie (4b) haben besondere Eigenschaften in Bezug auf die zulässigen logischen Schlussfolgerungen. Wenn wir beispielsweise (4a) und (4b) lesen, können wir folgern, dass die "Titanic" kein Boot, sondern ein Schiff, war.

Die vorliegende Dissertation verwendet ein von Smith (2003) eingeführtes Inventar von Situationstypen. Dieses schließt sowohl alle oben erwähnten relevanten aspektuellen Unterscheidungen als auch die Identifikation von generischen Ausdrücken mit ein. Situationstypen sind semantische Konzepte, die anhand ihrer internen temporalen Eigenschaften unterschieden werden (Smith, 2003, Seite 68). Es handelt sich bei Situationstypen um "verdeckte Kategorien" im Sinne von Whorf (1945), die jedoch mit verschiedenen linguistischen Formen korrelieren. Situationen werden in einem Diskurs von der *Verbkonstellation* eines Teilsatzes eingeführt, d.h. vom Hauptverb des Teilsatzes, dessen Argumenten und zugehörigen Adjektiven, Adverbialbestimmungen und Partikeln. Die Bestimmung des Situationstyps eines Teilsatzes erfordert daher die Kombination von lexikalischen und syntaktischen Faktoren und, in bestimmten Fällen, darüber hinaus Information aus dem Diskurskontext. Im Folgenden geben wir einen kurzen Überblick über Smiths Inventar von Situationstypen.

Zustände (States) sind Situationen, die für einen gewissen Zeitraum zutreffen wie in Beispiel (5). *Ereignisse* (Events) hingegen dauern eine gewisse Zeit oder geschehen zu einem bestimmten Zeitpunkt wie in (6). Events benutzen *dynamische* Verbkonstellationen, die andeuten, dass etwas geschieht; im Gegensatz dazu sind Zustände *statisch* und berichten über gleichbleibende Eigenschaften (Vendler, 1957).

(5) Der Graf besitzt den Bauernhof. (*Zustand*)

(6) John gewann das Rennen. (*Ereignis*)

Der Situationstyp *Generischer Satz* (Generic Sentence) beschreibt Fälle, in denen das Subjekt eines Satzes sich auf eine natürliche Gattung bezieht wie in (7). Sowohl (7b) als auch (8) sind habituell, da in beiden Fällen über Situationen verallgemeinert wird. In (7b) bezieht sich das Subjekt zusätzlich auf eine natürliche Gattung und der Satz wird daher als Generic Sentence eingestuft. Wenn das Subjekt ein bestimmtes Objekt referenziert, werden habituelle Sätze als *Generalisierender Satz* (Generalizing Sentence) markiert.

(7) (a) Löwen sind Fleischfresser. (*Generischer Satz*)
(b) Löwen fressen Fleisch. (*Generischer Satz*)

(8) Mary fährt mit dem Bus zur Arbeit. (*Generalisierender Satz*)

Smith führt außerdem zwei Situationstypen ein, die *Abstrakte Entitäten* beschreiben: *Fakt* (Fact) wird angewandt auf Teilsätze, die von Verben des Wissens eingebettet werden; *Behauptung* (Proposition) umfasst Fälle, in denen ein Gliedsatz eines Verb eingebettet wird, das Glauben oder Erwartung ausdrückt. In beiden folgenden Beispielen ist der unterstrichene Teilsatz derjenige, der als Abstrakte Entität verstanden wird.

(9) Ich weiß, dass Mary das Angebot nicht angenommen hat. (*Fakt*)

(10) Ich glaube, dass Mary das Angebot nicht angenommen hat. (*Behauptung*)

In dieser Arbeit werden zusätzlich zu den oben beschriebenen und von Smith eingeführten Situationstypen zwei weitere Typen verwendet. Es handelt sich dabei um *Frage* (Question) und *Imperativ* (Imperative), siehe Beispiele (11) und (12). Palmer et al.

(2007) erweiterten das Inventar um diese beiden Typen, um eine vollständige automatische Analyse der Sätze eines Textes zu ermöglichen. Keiner der anderen Situationstypen ist auf diese beiden Satzmodi anwendbar.

(11) Hat Mary das Angebot angenommen? (*Frage*)

(12) Komm bitte nicht so spät nach Hause! (*Imperativ*)

Palmer et al. (2007) erreichten in ihrem Annotationsprojekt für Situationstypen nur eine moderate Übereinstimmung zwischen den Annotatoren. Eine Ursache dafür war das Fehlen schriftlicher Annotationsrichtlinien. In dieser Arbeit wird aufgezeigt, dass Smiths Situationstypen entlang dreier Dimensionen unterschieden werden können, und dass die Vermittlung dieser einzelnen Unterscheidungen in der Summe zu einer konsistenteren Annotation von Situationstypen führt:

- Die **Aktionsart** des Hauptverbs eines Teilsatzes beschreibt auf Wortbedeutungsebene, ob es sich um ein *statisches* oder *dynamisches* Verb handelt. Dies zu erkennen ist wichtig für die Unterscheidung von *Zuständen* und *Ereignissen*.

- Ein weiteres Attribut (**Habituativ**) bezieht sich nicht auf die Wortbedeutung des Hauptverbs, sondern auf den gesamten Teilsatz. Ereignisse, die ein tatsächliches Geschehen ausdrücken, sind *episodisch*. Generische und generalisierende Sätze, die über Ereignisse generalisieren, sind *habituell*. *Zustände* und *Generische Sätze*, die keine Zustandsänderung implizieren, werden als *statisch* markiert.

- *Generische Sätze* machen eine Aussage über eine natürliche Gattung. Im Englischen ist in den meisten Fällen das Subjekt das *Topik* eines Satzes, d.h. der Teil des Satzes, über den eine Aussage gemacht wird. Daher werden in unserem Ansatz Subjekte als *generisch* markiert, wenn sie auf eine natürliche Gattung, oder auf eine beliebige Instanz der Gattung referenzieren.

Im ersten empirischen Teil der Arbeit wird ein umfassendes Korpus aus Texten verschiedener Genres und Domänen mit den obigen aspektuellen und semantischen Unterscheidungen und Situationstypen annotiert. Das Korpus setzt sich aus ca. 30000 Teilsätzen aus dem MASC-Korpus (Ide et al., 2008, 2010) und ca. 10000 Teilsätzen aus Wikipedia-Artikeln zusammen. Bei der Annotation der Situationstypen wird eine substanzielle Übereinstimmung erreicht. Es zeigen sich jedoch Unterschiede im Schwierigkeitsgrad, so sind etwa *Ereignisse* relativ leicht zu identifizieren, während es vergleichsweise schwieriger ist, generische Ausdrücke zu erkennen. Letzteres ist auch abhängig vom Genre: In enzyklopädischen Texten werden Nominalphrasen, die sich auf Arten beziehen, mit relativ großer Übereinstimmung annotiert, während in argumentativen Texten, die sich häufiger auf abstraktere Konzepte beziehen, mehr Schwierigkeiten auftreten.

Der zweite empirische Teil der Arbeit beschreibt, wie aus den Korpusdaten mit Hilfe von Algorithmen des maschinellen Lernens computergestützte Modelle für alle vier Klassifikationsaufgaben erstellt werden können. Diese werden sowohl quantitativ als auch qualitativ evaluiert. Für die Modellierung werden auf Entscheidungsbäumen basierende *Random Forests* (Breiman, 2001) und *Conditional Random Fields* (Lafferty et al., 2001) verwendet. *Conditional Random Fields* haben die Eigenschaft, dass die Klassifizierung der

Elemente einer Sequenz, die in unserem Fall aus den Teilsätzen eines Textes besteht, nicht unabhängig voneinander, sondern global optimiert getätigt wird. Auf diese Weise kann in nachweisbarem Maß Diskursinformation genutzt werden, um generische Ausdrücke zu identifizieren. Die Arbeit zeigt, dass die zu klassifizierenden Instanzen, d.h die einzelnen Teilsätze eines Textes, effektiv mit Hilfe von folgenden Attributen repräsentiert werden können:

- Die **Wortarten** der in einem Teilsatz vorkommenden Wörter sind hilfreiche Hinweise auf den Situationstypen. Sie spiegeln zu einem gewissen Grad wider, welches Tempus der Satz verwendet und ob Adverbien vorkommen.

- **Lexikalische Information** ist, vor allem wenn Trainingsdaten aus derselben Domäne vorhanden sind, äußerst wertvoll. Hier wird **distributionelle** Information verwendet, indem die Brown-Cluster-Identifikatoren (Brown et al., 1992) der in einem Teilsatz vorkommenden Wörter als Attribute verwendet werden.

- **Linguistische Indikatoren** (Siegel und McKeown, 2000) sind statistische Informationen, die die Verwendung eines Verbtyps in einem großen Textkorpus beschreiben. Es wird zum Beispiel erfasst, in wie viel Prozent des Auftretens eines Verbs dieses in der Verlaufsform (dem englischen *Progressive*) steht.

- **Syntaktisch-semantische** Attribute beschreiben das Hauptverb jedes Teilsatzes sowie dessen Subjekt. Die Attribute beinhalten unter anderem Tempus, Genus Verbi (aktiv oder passiv), WordNet-basierte Attribute (Fellbaum, 1998), grammatischen Aspekt (*Progressive* / *Perfect*), Dependenzrelationen, Artikel, Numerus und Person. Zusätzlich werden Attribute benutzt, die weitere Eigenschaften des Teilsatzes wie das Vorkommen von Negation, Konditionalen oder Modalverben erfassen.

Diese Arbeit zeigt, dass linguistische Indikatoren die **Aktionsart** eines Verbs als *statisch* oder *dynamisch* einordnen können, auch wenn in den Trainingsdaten nur ähnliche Verben vorkommen. Der Ansatz von Siegel und McKeown (2000) benutzt ausschließlich diese Attribute, was dazu führt, dass den Instanzen eines Verbtyps in den Testdaten immer dieselbe Klasse zugeordnet wird. Für in den Trainingsdaten vorkommende Verbtypen kann dies jedoch immer nur so gut funktionieren wie die Methode, jedem Verb einfach die jeweilige im Trainingskorpus vorkommende Mehrheitsklasse zuzuordnen. Im Fall von Verben, die mehrere Bedeutungen haben, von denen einige statisch und andere dynamisch sind, kann Kontextinformation aus dem Teilsatz in Form von syntaktisch-semantischen Attributen genutzt werden, um das System zu verbessern (siehe auch Friedrich und Palmer, 2014a). Die Genauigkeit des Modells liegt bei 80-90% und übertrifft den einfachen mehrheitsbasierten Ansatz in den Fällen, in denen ein Verbtyp im Korpus sowohl as *statisch* als auch als *dynamisch* vorkommt.

Im nächsten Schritt wird eine Methode vorgestellt, die eine vollständige automatische aspektuelle Klassifikation der Teilsätze eines Textes in die drei Klassen *episodisch*, *habitual* und *statisch* ermöglicht (Friedrich und Pinkal, 2015b). Eine vorherige verwandte Arbeit von Mathew und Katz (2009) basiert hingegen auf einer ausgewählten Zusammenstellung von Sätzen mit dynamischem Hauptverb und manuell erstellter syntaktischer Information. In der in diesem Experiment adressierten Klassifikationsaufgabe sind sowohl linguistische Indikatoren als auch syntaktisch-semantische Attribute notwendig, um eine gute Performanz zu erreichen. Die besten Ergebnisse werden erreicht, wenn man zunächst ein Modell trainiert und anwendet, das *statische* Fälle herausfiltert, und anschließend den übrigen Teilsätzen die Klassen *habituell* oder *episodisch* zuweist. Der gestufte Ansatz schneidet vor allem bei der schwierigen, da seltenen, Klasse *habituell* besser ab als ein Modell, das versucht, alle drei Klassen auf einmal zu unterscheiden.

Das dritte Experiment adressiert die automatische Erkennung von generischen Ausdrücken mit einem Fokus auf der Identifikation von Subjekten, die eine natürliche Gattung referenzieren. Es wird gezeigt, dass Diskursinformation durch das gleichzeitige automatische Annotieren der gesamten Teilsätze eines Textes mit Hilfe von *Conditional Random Fields* effektiv genutzt werden kann (siehe auch Friedrich und Pinkal, 2015a). Der hier vorgestellte Ansatz ist der erste, der eine Unterscheidung vornimmt zwischen Sätzen, die eine generische Aussage machen ("Der Blobfisch ist ein hässliches Tier"), und Sätzen, die eine Aussage über ein bestimmtes Ereignis mit Bezug auf eine natürliche Gattung machen ("Im September 2013 wurde der Blobfisch zum hässlichsten Tier der Welt gewählt"). Das beste hier präsentierte Modell erreicht eine Genauigkeit von 77,4% auf den Wikipedia-Daten. Ein Vergleichsmodell ohne Diskursinformation erreicht 74,0% und die Mehrheitsklasse im Datensatz beläuft sich auf 50,4%.

Basierend auf oben beschriebenen Ergebnissen präsentiert der letzte Teil der Experimente ein Modell, das wiederum mit Hilfe von Conditional Random Fields den Teilsätzen eines Textes Situationstypen zuweist (siehe auch Friedrich et al., 2016). Die Evaluation demonstriert, dass das Modell, welches alle oben beschriebenen Attribute verwendet, robuste Ergebnisse auch über Genre- und Domänengrenzen hinweg liefert. Das beste Modell annotiert 76,4% der Teilsätze mit dem richtigen Situationstypen. Die Mehrheitsklasse im Datensatz beläuft sich auf 45,0% und die mithilfe der manuellen Annotation ermittelte Obergrenze beträgt 79,6%.

Die in der vorliegenden Dissertation vorgenommene Studie und Modellierung von Situationstypen kombiniert eine Vielzahl von linguistischen Phänomenen an der Schnittstelle zwischen Syntax und Semantik und eröffnet daher neue Forschungsmöglichkeiten in verschiedene Richtungen. Aspektuelle Information ist relevant für die temporale Analyse von Texten, das Erkennen von Ereignissen in Abgrenzung zu generischer Hintergrundinformation in Form von Habitualen sowie für die Maschinelle Übersetzung. Die automatische Erkennung von Nominalphrasen, die sich auf eine Gattung beziehen, ermöglicht eine präzisere Extraktion von Informationen aus freiem Text und eine genauere automatische Auflösung von Koreferenzen. Die in dieser Arbeit untersuchten und modellierten linguistischen Phänomene beinhalten die Unterscheidung nach Situationstypen und reflektieren, wie der Autor oder Sprecher eine Situation in einem Diskurskontext repräsentiert. Diesen Teil der Bedeutung eines Textes zu modellieren, stellt einen wichtigen Schritt auf dem Weg zu automatischem Textverstehen dar.

Acknowledgments

During the years in which I worked on this thesis, I received invaluable support of several people. I am highly grateful towards my advisor Manfred Pinkal, who gently guided me from writing my first paper to becoming an independent researcher; for teaching me to always ask the questions "why?" and "what did we learn?"; for countless helpful discussions on semantic phenomena that I wanted to understand; for encouraging me to follow my own research interests; for always believing in me; and last, but not least, for his detailed and valuable feedback on all of my publications, including this thesis. I am equally grateful to my second advisor Alexis Palmer. She introduced me to the work of Carlota Smith and infected me with her enthusiasm for linguistics and the research area of aspect and discourse. She taught me writing, and so much more. This thesis would not exist without her, and her proofreading was, as always, indispensable. I would also like to thank Josef van Genabith for his interest in my work, his encouragement and for agreeing to be part of my committee.

I am grateful for the inspiring discussions I had during my internship in Edinburgh, especially with Bonnie Webber, Mark Steedman, Mike Lewis and Omri Abend. I also thank Nianwen Xue for our discussions on comparing the aspectual forms of English and Chinese, and all the brainstorming of where this research on aspect could eventually lead. These brainstormings are in part responsible for the ideas presented in Section 10.1.

Thanks to our student assistant Melissa Peate Sørensen for her dedication to this project. Our joint discussions contributed a lot to making our annotation guidelines clearer, and she helped with collecting and categorizing the Wikipedia part of our corpus (see Section 6.1).

I was lucky to have many great colleagues at Saarland University. Andrea Horbach provided valuable feedback during countless discussions, proofread almost all of my papers and this thesis, and kept my office plants alive. Ashutosh Modi was very helpful with some of the technical issues and Michaela Regneri provided me with lots of useful information in general (and coffee). For everything else, there was Diana Steffen. I also thank Stefan Thater (who is a great co-teacher), Vera Demberg, Ines Rehbein, Caroline Sporleder, Fatemeh Torabi Asr, Asad Sayeed and Alessandra Zarcone for many useful discussions related to the work presented in this thesis, and Hannah Kermes for proofreading the German part of this thesis.

I thank our annotators Ambika Kirkland, Fernando Ardente, Ruth Kühn, Melissa Peate Sørensen, Damyana Gateva, Kleio-Isidora Mavridou and Christine Bocionek for their extraordinary interest in this project and their very helpful observations and suggestions. Kleio-Isidora Mavridou and Liesa Heuschkel both did their master's theses in relation to this project, and I am very grateful for their interest and hard work. I also thank all the other students that I worked with on related projects during the time of my PhD studies, especially Jonathan Oberländer, whose implementation of a graphical parse visualization tool saved me a lot of time. I learned a lot from working with all of you, and I had a lot of fun.

Thanks also to Anna Nedoluzhko and Michal Novák, for the useful discussions during my

visit to Prague and their general support in many questions. Thanks to Roman Klinger for his helpful comments on the modeling part, especially for his suggestion to conduct the qualitative comparison presented in Section 8.4.4.

This research was supported by Cluster of Excellence on "Multimodal Computing and Interaction" (MMCI) of the German Research Foundation (DFG). I was also supported by an IBM PhD Fellowship. I thank the administrative staff of MMCI, Ursula Kröner and the Coli library staff for making my life so much easier.

I want to thank all my friends who were there for me in this time of my life, especially Casey Redd Kennington, with whom I share not only a great friendship but also a lively and fruitful scientific exchange, and who is always volunteering to proofread my English writings at short notice whenever I need it. Finally, my parents, Dominik, my grandma and the rest of my family – thank you for your never-ending love and support.

Contents

Part I

Introduction and background

Chapter 1

Introduction

When we use language to convey information, we locate the situations that we are talking about in time, i.e., in the past, present or future. To signal temporal locations or relations, many languages make use of *tense*, which amounts to choosing appropriate verb forms. In addition, language has means to present situations, which we assume to be expressed by the clauses of a discourse, in various *aspectual* manners. Aspect is a subsystem of language that represents situation from various viewpoints (Comrie, 1976; Smith, 1997). Distinctions that have been made in the linguistic and semantic theory literature include the classification of *states*, *events* and *processes* (Vendler, 1957; Bach, 1986).

This thesis addresses the computational processing of aspectual properties of clauses in a text. All examples in (1) are linguistic realizations referring to the same event, but focusing on different phases (Smith, 1997). The first clause presents the event in its entirety, while (b) and (c) "make visible" only an intermediate phase of the event. In the case of (c), the real-life "event" is even presented linguistically as a state.

(1) (a) The ship moved. (entire event)
(b) The ship was moving. (ongoing event / process)
(c) The ship was in motion. (state)

In addition to representing situations as one of the above situation types, we also express whether a situation regards a single event (2a) or whether we report a regularity. So-called *generic* clauses may report regularities generalizing either over events (2b) or, as in (2c), members of a kind (Krifka et al., 1995).

(2) (a) He went cycling yesterday.
(b) He cycles to work.
(c) Students like coffee.

This thesis work presents methods for automatically identifying the different types of situations expressed by the clauses of a text. This aspectual classification aims to identify the type of *linguistic representation* of particular situation occurrences in the world rather than to classify the occurrences themselves (Filip, 2012, sec. 2.5.6). We classify the linguistic representation of the event at the clause level, distinguishing between the different cases given in (1a-c). The task addressed here is orthogonal to work on *semantic parsing*, where the central interest is the identification of predicate-argument structures. The latter, answering the question "who did what to whom" would represent the event in (1) for example as `move(agent:ship)`. What semantic parsing and aspectual classification have in common is that both target representing meaning in a manner that facilitiates automatic text understanding. There is extensive recent work on semantic parsing using various formalisms including PropBank (Palmer et al., 2005), FrameNet (Ruppenhofer et al., 2006) and Abstract Meaning Representations (Banarescu et al., 2013). However, there is relatively little work in computational linguistics that addresses aspectual classification. In order to understand textual discourse in the sense of correctly interpreting temporal relations and distinguishing specific from more general information, the two tasks are both necessary and complementary. Consider the effect of using sentences (a), (b) or (c) of example (1) before uttering "She waved goodbye." If used after (b) or (c), we infer that the waving happened during the ship's moving, but if used after (a), we arrive at the somewhat odd interpretation that the ship moved, stopped, and that only then she waved. This example illustrates how we use aspect to infer temporal relations between situations. In tenseless languages such as Mandarin Chinese, this is actually one of the predominant mechanisms for arriving at temporal interpretations (Smith and Erbaugh, 2005).

This thesis also addresses the recognition of two phenomena related to *genericity* (see Section 2.3). The first phenomenon are *habituals*, which are sentences that generalize over situations such as (3b). In contrast, (3a) reports on a single event. Habituals allow exceptions: we still consider (3b) to be true if the ship has not been painted in a particular year, but in most other years.

(3) (a) They painted the ship. (one-time event)
(b) They paint the ship once a year. (regularity: habitual)

The second phenomenon is *reference to kinds* as in (4b), which is in contrast with sentences making statements about particular object (4a).

(4) (a) The RMS Titanic was a British passenger liner. (object-referring)
(b) Ships are generally distinguished from boats based on size, shape and cargo or passenger capacity. (kind-referring)

Generic sentences such as (4b), which make statements about kinds, have special properties regarding the logical inferences that they allow. For instance, upon hearing (4a) and (4b), we can infer that the "Titanic" was a ship, not a boat.

In this thesis, we adopt the inventory of *situation entity types* as proposed by Smith (2003). It has the advantage of addressing all the above mentioned relevant aspectual distinctions, including the identification of stative versus eventive clauses as well as genericity. We introduce details of this inventory in Section 1.1. Previous related work in computational linguistics using this inventory consists of a domain-dependent system trained on a relatively small data set (Palmer et al., 2007). Other related works on creating automatic systems for the aspectual distinctions addressed in this thesis only address parts of the relevant phenomena, i.e., genericity (Reiter and Frank, 2010), inherent lexical aspect (Siegel and McKeown, 2000) and habituals (Mathew and Katz, 2009). These works mostly use manually selected sets of sentences for their studies.

The two major contributions of this thesis are (a) an in-depth corpus-linguistic study on situation entity types using full-text annotation, and (b) the development of computational models for automatically predicting situation entity types, lexical aspectual class and habituality for clauses as well as genericity for their subjects.

We develop an annotation scheme that leads to a corpus annotated with substantial agreement and conduct a quantitative and qualitative analysis of differences in situation entity type distribution across 13 genres. To make the annotation task more transparent and enable a fine-grained analysis, we collect labels not only for situation entity types but also for relevant sub-tasks. In other words, we work in the framework of Smith's (2003) situation entity types in order to address three related aspectual distinctions, as well as the overall situation entity type classification task on full texts. Specifically, we label the main verb of each clause as being stative or dynamic at the word-sense level; we mark clauses as habitual, episodic or stative, and we mark the subject of each clause as a reference to a particular individual or a kind. Using the resulting data set, we create computational models for situation entity types, as well as classifiers addressing the related sub-tasks.

Our computational models are trained and evaluated on a multi-genre corpus of approximately 40,000 clauses from MASC (Ide et al., 2008) and Wikipedia which have been annotated with substantial agreement. We train and test our models both within genres and across genres, highlighting differences between genres and creating models that are robust across genres. Both the corpus and the code for an situation entity type labeler are freely available.[1] These form the basis for future research on situation entity types and related aspectual phenomena and will enable the inclusion of situation entity type information as a preprocessing step into various natural language processing tasks. For instance, information extraction or temporal relation processing are expected to profit from accurate aspectual classification of clauses (Van Durme, 2009; Costa and Branco, 2012). We discuss the potential of aspectual classification for natural language processing further in Chapter 10.

Section 1.1 of this chapter introduces the inventory of situation entity types that we are working with; Section 1.2 gives an overview of the structure of this thesis; Section 1.3 lists our contributions and related publications.

1.1 Situation entity types

The inventory of situation entity types that we adopt in this thesis was introduced by Smith (2003, p. 1) in her work on *discourse modes*, which are "classes of discourse passages, defined by the entities they introduce to the universe of discourse and their principle of

[1] www.coli.uni-saarland.de/projects/sitent

Situation entity type	Description	Example
Eventualities		
State	introduce properties	The colonel owns the farm.
Event	happenings	John won the race.
Report	for attribution	“...”, said Obama.
General Statives		
Generic Sentence	generalizations over kinds	The lion has a bushy tail.
Generalizing Sentence	habituals: generalizations over situations	Mary often fed the cat last year.
Abstract Entities		
Fact	clausal complements of verbs of knowledge	I know that she refused the offer.
Proposition	clausal complements of verbs of belief	I believe that she refused the offer.
Question		Who wants to come?
Imperative		Hand me the pen!

Table 1.1: Inventory of situation entity types, adapted from Smith (2003) and Palmer et al. (2007).

progression.” This thesis addresses the automatic classification of the aspectual entities that make up the first part of Smith’s definition. An overview of the inventory is given in Table 1.1. Situation entity types are “semantic concepts organized according to their internal temporal properties” (Smith, 2003, p. 68). They are covert linguistic categories in the sense of Whorf (1945), tacitly available to the speakers of a language, and they have linguistic correlates. Situation entities are introduced to the discourse by a clause’s *verb constellation*, i.e., the clause’s main verb and some or all of its arguments and modifiers (for more details see Chapter 4). Deciding on the type of a situation entity thus involves the combination of lexical and syntactic factors, as well as, as we will show, information from the surrounding discourse.

Smith’s (2001, 2003, 2005) situation entity types include the “particular” situation types State (5) and Event (6). States are situations that hold in time, while Events take place in time. Events use *dynamic* verb constellations and report that something happened; States are generally *static* and introduce properties to the discourse (Vendler, 1957; Bach, 1986).

(5) The colonel owns the farm. (State)

(6) John won the race. (Event)

In English, linguistic correlates for States and Events include, for example, their use with the Progressive (Smith, 2003, p. 75). Events are grammatical in the Progressive (7), while STATES as in (8) are not.[2,3]

(7) John was washing the car. (EVENT)

(8) *Mary was knowing the answer. (STATE)

General Statives include GENERIC SENTENCES (9), which hold of kinds, and GENERALIZING SENTENCES (10), which invoke patterns of situations. Smith (2003) refers to Krifka et al.'s (1995) seminal article on genericity. General Statives are therein also called *characterizing sentences*. They are not marked by particular morphemes or verb class (Smith, 2003, p. 77), but distributional characteristics that distinguish them are the occurrence of kind-referring noun phrases or the fact that a frequency adverb such as "usually" or "typically" can felicitously be added to the sentence.

(9) The lion has a bushy tail. (GENERIC SENTENCE)

(10) Mary often fed the cat last year. (GENERALIZING SENTENCE)

Abstract Entities describe types of situations that are not spatiotemporally located in the world, rather, they are *about* the world (see also Asher (1993)). Smith introduces the two sub-types FACT (11), which introduces assessments of abstract or concrete states of affairs, and PROPOSITION (12), which are the objects of mental states such as beliefs, expectations and decisions. In each case, the Abstract Entity is the underlined part of the example.

(11) I know <u>that Mary refused the offer</u>. (FACT)

(12) I believe <u>that Mary refused the offer</u>. (PROPOSITION)

These situation entities are expressed as clausal arguments of certain predicates such as (canonically) "know," "realize" or "believe," which, in turn, usually introduce STATES or EVENTS to the discourse. Following Smith (2003), we use PROPOSITION in a different sense than the usual meaning of "proposition" in semantics - naturally situation entities of any type may have propositional content. Smith's use of the term (and thus ours too) contrasts PROPOSITION with FACT – our PROPOSITIONS are simply sentences presented as a belief of the writer or speaker, regardless of whether their propositional content is true or not. This use of "proposition" also occurs in linguistic work by Peterson (1997) on factive versus propositional predicates.

Theoretically speaking, situation entity types are a "closed system" (Smith, 2003, p. 68), i.e., a choice must be made from a few possibilities. In practice, however, cases are not so clear-cut. A first corpus creation project (Palmer et al., 2007) based on intuitive annotation resulted only in moderate inter-annotator agreement. In this thesis work, we thoroughly investigate the conceptual levels and factors that are relevant when differentiating between Smith's situation entity types. We identify three relevant dimensions along which situation entity types can be distinguished, and which help annotators to consistently label clauses with their type:

[2] In examples, * indicates ungrammaticality.

[3] Following Comrie (1976), we use initial capitals for the names of language-particular categories and lower case for language-independent semantic distinctions.

- The **lexical aspectual class** of a clause's main verb describes, at the word-sense level, whether it is ***stative*** or ***dynamic***. Recognizing this feature is necessary to distinguish States from Events.

- In contrast, **habituality** is a feature at the clause level. Events that happen are ***episodic***, Generalizing Sentences or Generic Sentences that refer to patterns of situations are ***habitual***, and States or Generic Sentences that do not involve change are ***static***.

- Finally, Generic Sentences are defined as clauses making statements about kinds; the main verb's subject (which is, in English, usually the clause's topic) is marked as ***generic*** if it refers to a kind or an arbitrary member of a class.

There are three additional situation entity types – Report (for attribution), Question and Imperative – which have been introduced by Palmer et al. (2007) in order to be able to mark texts exhaustively. Clauses belonging to one of these three types do not fit into any of the above categories. Chapter 5 explains our annotation guidelines, giving more details on our own implementation of Smith's inventory and the three related features described above.

The different computational modeling tasks addressed here all require the representation of clauses using lexical and syntactic attributes for the clause's verb constellation including clausal modifiers. A good predictor for lexical aspectual class of a verb type is its behavior in a large text corpus, i.e., how often it occurs with linguistic indicators such as the Progressive (Siegel and McKeown, 2000). In addition, the arguments of the verb in context and its grammatical tense and aspect must be taken into account, as illustrated by (13).

(13) (a) Water fills the pool. (***stative***)
(b) She filled the glass with juice. (***dynamic***)

While it is sufficient to model lexical aspectual class at the level of single clauses, the interpretation of habituality, genericity and hence also situation entity types requires knowledge about a clause's discourse context. Example (14), from Mathew and Katz (2009), illustrates that determining whether the respective second sentence is habitual is not possible without its discourse context.

(14) (a) John rarely ate fruit. He just ate oranges.
(b) John didn't eat much at breakfast. He just ate oranges.

The work presented in this thesis uses English texts; however, while our annotation guidelines are specific to the English language, the aspectual distinctions and the inventory of situation entity types are not. They can in principle be implemented for other languages as well (Mavridou et al., 2015).

1.2 Plan of this thesis

Part I: Introduction and overview

The first part of this thesis sets the background for our corpus-linguistic and empirical work on aspectual clause types and genericity. **Chapter 2** gives an overview of the related linguistic phenomena in the linguistic and semantic theory literature. **Chapter 3** reports on the respective related work in computational linguistics.

Part II: Corpus: annotation and agreement

The second part describes the construction of our multi-genre corpus annotated with situation entity types. The first question we address (**Chapter 4**) is how to segment texts into units that receive a situation entity type label. We link this question to previous research on discourse segmentation, and explain our heuristic method for identifying relevant segments based on an existing discourse parser.

Chapter 5 explains our annotation scheme and guidelines for the situation-related features of lexical aspectual class, genericity and habituality, as well as the full inventory of situation entity types. **Chapter 6** describes our corpus of Wikipedia and MASC (Ide et al., 2008) texts including the development of our guidelines. With the aim of determining the difficulty of making explicit the aspectual distinctions, which native speakers use and interpret subconsciously with great ease, we analyze inter-annotator agreement, bias of individual annotators as well as intra-annotator agreement, i.e., how consistently annotators reproduce their own decisions. We describe the construction of our gold standard based on majority voting and the resulting label distributions in the various genres of our corpus. Finally, we discuss difficult cases.

Part III: Methods and experimental evaluation

In this part, we describe our methods for creating computational models of aspect as well as their evaluation. We represent each clause using a variety of syntactic-semantic features and conduct our experiments in a supervised classification setting using Random Forest classifiers, maximum entropy models and conditional random fields (**Chapter 7**). **Chapter 8** then reports the individual set-ups and results of our experiments on classifying the lexical aspectual class of each clause's main verb, determining whether a clause is habitual, episodic or static, and whether the subject of a clause refers to a kind. The findings from the experiments on these sub-task allow us to create a situation entity classifier which is not only robust across genres but which also performs well when comparing to the upper bound established by measuring human performance on the task.

Part IV: Future directions and conclusion

The focus and scope of this thesis is a corpus-linguistic approach to the linguistic and semantic phenomena of aspect and genericity as well as their computational modeling. In

Chapter 9, based on a corpus study, we discuss how these phenomena – and specifically situation entity types – relate to temporal information in discourse and how automatic temporal relation processing could profit from including them as a source of information. We outline additional directions for future work in **Chapter 10** and summarize in **Chapter 11**.

1.3 Contributions of this thesis

The main contributions of this thesis consist of a large corpus-linguistic analysis of clause-level aspect and genericity as well as their computational modeling:

1. A large multi-genre corpus of English texts reliably annotated for situation entity types and related aspectual phenomena as well as genericity; including a cross-linguistic case study for German; joint work with Alexis Palmer, Manfred Pinkal, Melissa Peate Sørensen and Kleio-Isidora Mavridou (Friedrich and Palmer, 2014b; Friedrich et al., 2015b; Mavridou et al., 2015).

2. A scalable approach for labeling verbs *in context* with their fundamental lexical aspectual class as ***stative*** or ***dynamic***; joint work with Alexis Palmer (Friedrich and Palmer, 2014a).

3. A novel sequence labeling method for identifying generic expressions which successfully leverages discourse context and carefully distinguishes between noun-phrase level and clause level genericity; joint work with Manfred Pinkal (Friedrich and Pinkal, 2015a).

4. Bringing together two strands of previous work that only address the stative-dynamic and episodic-habitual distinctions, respectively, as a consequence creating the first fully automatic approach for classifying *all* clauses of a text with respect to their clause-level aspectual nature; joint work with Manfred Pinkal (Friedrich and Pinkal, 2015b).

5. The first robust approach to automatically labeling clauses with their situation entity type, introducing the use of distributional information and including a detailed cross-genre evaluation; joint work with Alexis Palmer and Manfred Pinkal (Friedrich et al., 2016).

Conclusions. In our corpus study, we find that in general, situation entity types can be annotated with substantial agreement. However, some of the decisions are easier than others, e.g., it is relatively easy to identify Events, while annotators have more difficulties with recognizing genericity. This, in addition, also differs per genre, e.g., it is comparably easier to recognize kind-referring noun phrases in encyclopedic tests than in argumentative essays, which more often refer to abstract notions.

Our computational models for lexical aspectual class, habituality, genericity and situation entity types all improve upon the previous state-of-the-art. The reasons for this are (a) our use of extended feature sets capturing important details within clauses, (b) our use of highly powerful discriminative sequence labeling models, which capture the surrounding discourse context at a high level, and (c) our large corpus, which allows for a more detailed an principled study of the phenomena and the identification of the information relevant to the classification tasks. Depending on the task, we reach accuraries between 75% and 85%, which provides a performance level that facilitates future research using the predictions of our models as input to various natural language processing tasks.

Relevance for computational linguistics. There are various sub-tasks within computational linguistics that will potentially benefit from including aspectual information. As already illustrated in this introduction, temporal reasoning (Lascarides and Asher, 1993; Passonneau, 1988; Verhagen et al., 2007, 2010; UzZaman et al., 2013; Bethard et al., 2016), i.e., the interpretation of the temporal structure of texts, is a research area that will certainly profit from modeling the aspectual distinctions addressed in this thesis. Stative and dynamic clauses lead to different interpretations of how two consecutive clauses are related temporally (see our discussion of example (1)); identifying generalizations, which provide background information, is relevant as they should not usually be linked into the story's foreground. Information extraction (Sarawagi, 2008; Jiang, 2012) and factuality recognition (de Marneffe et al., 2012) are also areas that will profit from accurate automatic identification of clauses that express particular situations versus clauses that are habituals. Especially for information extraction, but also for automatic coreference resolution, it is important to correctly distinguish mentions referring to particular individuals from kind-referring mentions. More details and motivating examples regarding potential applications are given in Section 10.1.

Relevant publications

The following publications report on parts of the research described in this thesis:

Annemarie Friedrich and Alexis Palmer. Situation entity annotation. In *Proceedings of the 8th Linguistic Annotation Workshop (LAW VIII)*, Dublin, Ireland, August 2014b

Annemarie Friedrich and Alexis Palmer. Automatic prediction of aspectual class of verbs in context. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, Baltimore, USA, June 2014a

Annemarie Friedrich, Alexis Palmer, Melissa Peate Sørensen, and Manfred Pinkal. Annotating genericity: a survey, a scheme, and a corpus. In *Proceedings of the 9th Linguistic Annotation Workshop (LAW IX)*, Denver, Colorado, USA, June 2015b

Annemarie Friedrich and Manfred Pinkal. Discourse-sensitive Automatic Identification of Generic Expressions. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL)*, Beijing, China, July 2015a

Annemarie Friedrich and Manfred Pinkal. Automatic recognition of habituals: a three-way classification of clausal aspect. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Lisbon, Portugal, September 2015b

Annemarie Friedrich, Alexis Palmer, and Manfred Pinkal. Situation entity types: automatic classification of clause-level aspect. In *In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, Berlin, Germany, August 2016

Additionally, some preliminary results of ongoing work have been published in the following articles:

Alexis Palmer and Annemarie Friedrich. Genre distinctions and discourse modes: Text types differ in their situation type distributions. In *Proceedings of the Symposium on Frontiers and Connections between Argumentation Mining and Natural Language Processing*, Bertinoro, Italy, July 2014

Kleio-Isidora Mavridou, Annemarie Friedrich, Melissa Peate Sorensen, Alexis Palmer, and Manfred Pinkal. Linking discourse modes and situation entities in a cross-linguistic corpus study. In *Proceedings of Linking Models of Lexical, Sentential and Discourse-level Semantics (LSDSem 2015)*, Lisbon, Portugal, September 2015

Chapter 2

Related work in theoretical linguistics

In this chapter, we provide a survey of the linguistic and semantic theories which form the basis of the corpus-linguistic work presented in Part II of this thesis. As described in Chapter 1, we have adopted Smith's (2003) situation entity types as our inventory of clause types. Smith provides many examples, but she does not completely map out the relationships of the situation entity types with aspectual and semantic distinctions made elsewhere in the literature. This chapter gives an overview of the broader range of linguistic theories on aspect (Section 2.1 and Section 2.2) and genericity (Section 2.3), in the context of which Smith develops her situation entity types. The purpose of this chapter is to provide the necessary linguistic background for readers with a computational background, and to introduce the terminology used in this thesis. The mapping between the situation entity types and the phenomena described here is part of the contribution of this thesis and is described in Chapter 5 in detail.

Aspectual meaning (both eventuality type and grammatical aspect) refers to how the internal structure of situations is presented, and applies for sentences rather than for verbs or verb phrases. It is compositional, composed by the verb's inherent meaning, its arguments, tense, morphological aspect markers and the adverbs of the sentence (Verkuyl, 1972; Mourelatos, 1978; Smith, 1997). Aspectual distinctions are semantic in nature. They may or may not be grammaticalized or lexicalized in individual languages (Comrie, 1976, p. 7), i.e., they may be covert categories in the sense of Whorf (1945). Section 2.2.4 will elaborate on this. Regardless of whether or not they are grammaticalized or lexicalized, the categories presented here are assumed to exist in the semantics of a language (Verkuyl, 1972).

Comrie (1976) notes that in addition to the lack of a generally accepted linguistic terminology related to aspect, there is a terminological and conceptual confusion of tense and aspect. This is at least partially caused by the fact that the expression of temporal location is intertwined morphologically with aspect in many languages of the world (Smith, 1997, p. 97), so the traditional grammatical terminologies of individual languages often introduce tense/aspect combinations as *tenses*. While aspect refers to situation-internal time, tense, in contrast, is *deictic* and locates the time of a situation relative to the situation of the utterance or relative to another situation. Tense can thus be regarded as situation-external time.

Smith's Generic Sentences are defined as sentences that "refer to kinds rather than individuals," hence, Section 2.3 introduces the relevant linguistic notions of genericity. As Smith develops her situation entity types as the basis for her theory of discourse modes, at the end of this chapter we give a brief overview of how the aspectual notions presented here are linked with linguistic theories of discourse (Section 2.4).

2.1 Inherent lexical aspect

Comrie (1976), slightly generalizing from an earlier definition by Holt (1943), defines *aspects* as "different ways of viewing the internal temporal constituency of a situation," i.e., as ways of presenting situation-internal time. He uses *situation* as a cover term for states, events or processes, which corresponds to our use of Smith's term *situation entity* in the previous chapter. This chapter addresses categorizations that have been treated in the linguistic and semantic theory literature under the term *inherent lexical aspect* (Comrie, 1976), also called *situation type* (Smith, 1997), *eventuality type* (Bach, 1986), or *aktionsart* (e.g., Hinrichs, 1985).[1] This categorization is concerned with the lexical properties of verb senses in their context, including whether the eventuality type is stative or dynamic ("know" vs. "buy") and whether it includes a "natural endpoint" ("write a letter" vs. "sing"). It is important to note that not verb types, but verb usages or verbs in context or rather sentences have lexical aspectual class.[2] In Section 2.1.4 we will discuss various mechanisms which cause shifts from one aspectual type to another. Nevertheless, according to Smith (1997, p. 54), the verb is the aspectual center of a sentence, as verbs have an intrinsic aspectual value. This means that for verbs – in contrast to other parts-of-speech – it is possible to apply the aspectual distinctions described in this section without having additional context. For more details on lexical aspect, see the survey by Filip (2012).

2.1.1 Stativity

The most fundamental distinction made in the hierarchies of eventuality types presented in Section 2.1.3 below is the one between states ("love", "own") and events ("run", "buy"). In contrast to dynamic predicates, state verbs entail no change (Filip, 2012). States obtain in time but they do not take time; events occur, happen or take place (Smith, 1997).

Vendler (1957) suggests the "progressive test" to identify stative verbs in English, assuming that these are infelicitous with the progressive. However, Dowty (1979) notices that this test is insufficient as there are many stative verbs that may occur in the progressive, such as "lie" in (1).

(1) Socks are lying under the bed.

The progressive test only works for stative predicates that are *individual-level* predicates

[1] We use these terms interchangeably in this chapter, following those primarily used in the respective literature.

[2] The literature mostly speaks of *sentences* as S nodes; naturally, the different clauses of longer sentences will each have their own aspectual class.

in the sense of Carlson (1977b) – called *object-level* by Dowty – i.e., predicates that attribute properties to an individual for the entire duration of the individual's existence. For example, sentence (2) is ungrammatical because it is an object-level state according to Dowty (1979, p. 180).

(2) *John is knowing the answer.

Cases such as (1), which are felicitous in the progressive, are *stage-level* states in the sense of Carlson. Dowty calls verbs such as "sit," "stand," or "lie" *interval states*.

Another criterion along which situations can be distinguished is whether the phases of a situation differ or are all the same (Vendler, 1957; Smith, 1997; Dowty, 1986). States have the "subinterval property" (Bennett and Partee, 1978); roughly speaking, if a sentence is true in a time interval *I* and its main verb phrase is a *subinterval VP*, then the sentence is true at every subinterval of I. For example, if "John has blue eyes" is true for the interval of his life, then for each timespan within his life "John has blue eyes" is also true. In contrast, if we can say that "John wrote a letter from 5pm to 6pm," this entails that he finished the letter only at 6pm. Hence, it would be wrong to say "John wrote a letter from 5.15pm to 5.30pm." Saying that he was busy writing a letter, without stating that he finished doing so, requires the choice of the Progressive in English.

While states are generally continuous, they may have a beginning and / or an end (Comrie, 1976, p. 59). They are *durative*, i.e., conceived as lasting for a certain period of time. The opposite are situations that are perceived as punctual (Comrie, 1976, p. 41). Punctual situations are always dynamic as they involve a change of state by definition, as illustrated by example (3).

(3) John reached the goal.

A special class of punctual predicates are called *semelfactives* by Smith (1997). They include verbs such as "hiccup" or "blink." Semelfactive verbs return to their initial state at their end and they often occur as iteratives (Filip, 2012).

Another special case are perception verbs such as "hear" or "see"; they can often be classified either as a state or as an activity, i.e., an event that extends in time and that does not have a natural endpoint. Individual languages often make arbitrary choices of whether such verbs are classified as stative or not (Comrie, 1976, p. 35).[3] Sentences like (4) and (5) are both classified as stative by Vendler (1957).

(4) I saw a star from my window. (Vendler: *state*; Mourelatos/Smith: *state*)

(5) I saw him run. (Vendler: *state*; Mourelatos/Smith: *event/activity*)

Mourelatos (1978) argues that the latter example conveys that an event of "seeing" must have happened. Similarly, Smith (1997, pp. 56–57) observes that this example cannot simply be explained by assuming a "spotting" sense for "see," as the "seeing" extends in

[3] This fact may admittedly contribute to some disagreements in the annotations of our corpus as some of our annotators are non-native speakers of English.

time and should thus be regarded as an activity. Yet, the linguistic test for activities, which are normally compatible with the progressive, does not work for perception verbs, i.e., "I was seeing" is wrong. This behavior is unique for perception verbs.
Another difficult case are sentences like (6).

(6) I work for IBM.

Habits, including occupations, dispositions and abilities, are states in Vendler's (1957) sense. "Are you smoking?", in his account, asks about an activity, while the question "Do you smoke?" asks about a state. He calls these cases *derived* states. He applies the same analysis to cases such as (6).

2.1.2 Telicity

The second important distinction related to inherent lexical aspect is *telicity*, i.e., the inclusion or noninclusion of a goal in the lexical sense of a verb in a given context. The term *telic* was introduced by Garey (1957) and is derived from Greek *télos* (goal). In his definition, telic verb senses have a built-in goal: when the goal of a telic event is reached, a change of state occurs and the event is complete Smith (1997, p. 19). When a telic verb is used in the perfective, it means that the goal is reached at the time of reference. The imperfective applied to a telic verb "hides the arrival or nonarrival at the goal" (Garey, 1957). If (7a) is true at some particular point in time, it cannot be the case that (7b) is true at the same point in time.

(7) (a) John was recovering.
(b) John has recovered.

In contrast, *atelic* verbs are those which do not have to wait for a goal for their realization, they are realized as soon as they begin. They do not have an outcome; they are processes that can stop at any time. If an atelic verb is used in imperfective form (8a), we can infer that the sentence in perfective form (8b) is also true.

(8) (a) He was singing.
(b) He has sung.

A simple linguistic test for telicity in English is the combination with *in*-NP and *for*-NP modifiers as in example (9): only telic verbs combine with the former, while only atelic verbs combine with the latter (Vendler, 1957).[4]

(9) (a) John recovered in an hour / *for an hour. (*telic*)
(b) John swam *in an hour / for an hour. (*atelic*)

Mourelatos (1978) suggests that telic verbs are like count nouns, while atelic verbs are like mass nouns in the sense of "having natural endpoints." Count nouns "apple," "woman")

[4] Example (9) taken from Filip (2012).

Situation type	dynamic	durative	telic	examples
State	-	+		know the answer, love Mary
Activity	+	+	-	laugh, stroll in the park
Accomplishment	+	+	+	build a house, walk to school
Achievement	+	-	+	win a race, reach the top
Semelfactive	+	-	-	tap, knock

Table 2.1: Situation types (Vendler, 1957; Smith, 1997, p. 20).

take the indefinite article and cardinal numbers (e.g., "an apple," "three women"). Mass nouns ("hunger," "snow," "beer") do not generally have plural forms, or, if used in plural, they usually mean something different ("three beers" = "three types / glasses of beer"). Similarly, as illustrated by the sentences in (10), telic verbs can be combined with count adverbials but not vague quantifiers such as "a lot"; atelic verbs behave in the opposite way.

(10) (a) John cooked dinner three times / *a lot. (*telic*)
(b) John swam *three times / a lot. (*atelic*)

Comrie (1976, p. 44) requires that both telic and atelic predicates have processes that lead up to their built-in terminal point, i.e., he considers both situation types to be durative. He argues that sentences like (11) are atelic as "reaching" does not include the process leading up to it.

(11) John reached the summit.

Smith (Smith, 1997, p. 30) offers a different interpretation, noting that cases such as (11) may require preliminary stages, but they may consist of a process that is "detached" from the predicate itself. She defines telic predicates as having an intrinsic bound, which results in a change of state; consequently, she analyzes examples such as (11) as telic.

2.1.3 Eventuality types

With the two important distinctions of stativity and telicity established, we are now ready to explain some influential inventories of eventuality types. Table 2.1 shows the set of situation types introduced by Vendler (1957): *state*, *activity*, *accomplishment* and *achievement*. In his original account, they are taken to apply at the level of verb senses. He calls them time schemata for verbs, though he notes that one verb type can be used according to different schemata. Mourelatos (1978) criticizes that these earlier analyses (Vendler, 1957; Kenny, 1963) focus too much on predicates that require human agency, and suggests the terminology presented in Figure 2.1. *Performances* (Kenny, 1963) are "actions that tend towards a goal." Vendler divides them further into accomplishments

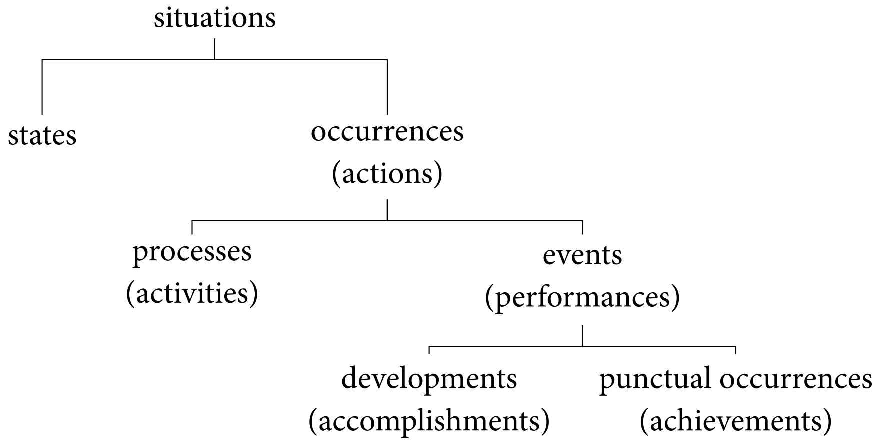

Figure 2.1: Classification of aspectual oppositions according to Mourelatos (1978), the terms used by Vendler (1957) and Kenny (1963) are shown in parentheses.

and achievements. We will use Vendler's terminology here, as it is the most widely used one, without assuming human agency for the predicates.

States are inherently stative and atelic, and they have neither different successive phases nor a predefined endpoint. They are durative as they usually extend in time, though interval states may be true only for very short periods of time (12).

(12) He was very proud of himself for just two seconds (*state*)
(then he realized his mistake).

The other three Vendlerian classes are all dynamic. They differ in whether they have a built-in endpoint, and in whether they have a clearly defined process leading up to this endpoint. Activities use atelic verbs, consisting entirely of a process. The successive phases that the process is composed of do not necessarily have to be completely homogenous, as (13) illustrates.

(13) He was breathing. (*activity*)

Accomplishments (14) have a different internal structure: they consist of a process that leads up to a built-in terminal point.

(14) He wrote a letter. (*accomplishment*)

Like accomplishments, achievements (15) have a terminal point including a change of state, but the verb meaning does not include a process leading up to this point.

(15) He arrived at the station. (*achievement*)

As indicated in Section 2.1.2, Comrie (1976) and Smith (1997) differ in their analyses of telicity, which, on the level of eventuality types, results in a different definition of achievements. Focusing on the change of state, Smith analyses them as instantaneous and telic; Comrie calls them atelic because the predicate does not include a process leading up to it.

	Events		**States**
	atomic	extended	
+ consequent state	**culmination** recognize, spot, win the race	**culminated process** build a house eat a sandwich	understand, love, know, resemble
- consequent state	**point** hiccup, tap, wink	**process** run, swim, walk play the piano	

Table 2.2: Eventuality types of Moens and Steedman (1988).

For "punctual" or single-stage events which do not cause a change of state, Smith (1997) adds the situation type *semelfactives*, which include verbs like "knock," "flash" or "blink." Semelfactives are interpreted iteratively when used with *for*-adverbials as in (16) (Filip, 2012). Both achievements and semelfactives are instantaneous, but in contrast to the latter, the former result in a change of state.

(16) The light was flashing for an hour.

2.1.4 Aspectual type coercion

In this section, we introduce the ideas of Moens and Steedman (1988) on aspectual *coercion*, which refers to the process by which aspectual types of verbs are shifted based on their arguments or other aspectual operators such as adverbials. Similarly, situation entity types can be *derived* from other types (Smith, 2003), e.g., negated Events are classified as States.

Moens and Steedman work with an inventory of eventuality types similar to the ones introduced above. They also distinguish states from events; the latter are "happenings with defined beginnings and ends" and include ongoing processes or activities. Moens and Steedman's types, aspectual profiles of sentences used in a context, are classified by making reference to a so-called *nucleus*, which consists of a preparatory process, a culmination point and a consequent state. The culmination point is a "goal event" at which a change into the consequent state happens. Sentence (17) is a typical example for a case that has all of these phases: a preparatory phase (the house is being built), a culmination (the moment at which it is completed) and a consequent state (the house is complete). Such cases are called *culminated process*, and correspond to accomplishments.

(17) John and Mary built a house.

Other eventuality types, as illustrated in Table 2.2, may either lack the preparatory process, i.e., they are *atomic*, or they may not include a consequent state. Example (18) only consists of a culmination: it is a punctual and instantaneous event, classified as an achievement above.

(18) John reached the top.

Punctual expressions without a result state – Smith's semelfactives – are called *point expressions* here. Activities consist only of a process; they do not have a culmination point. Moens and Steedman explicitly allow for shifts from one type to another due to aspectual operators. For example, the English Progressive signals an ongoing process. When used with a predicate whose lexical entry corresponds to a culminated process, its function is to strip off the culmination point and only make visible the preparatory process. This leads to an elegant resolution of the *imperfective paradox* (Dowty, 1979; Lascarides, 1991), the observation that from (19a) we cannot infer (19b).

(19) (a) John was running a mile.
(b) John ran a mile.
(c) John was running a mile, but he gave up after five minutes.

Moens and Steedman's analysis easily explains why (19c) is grammatical: (19a) makes visible only the preparatory phase of the predicate, coercing the sentence to a process without a culmination. Similarly, earlier accounts had trouble explaining cases like (20a), where an achievement predicate (a culmination) occurs in the Progressive, which signals an ongoing process. Moens and Steedman assume that by the principles of aspectual coercion, a preparatory process is added, resulting in a culminated process. Then the culmination point is stripped off, again indicating a process which does not necessarily result in a change of state – here it would be grammatical to utter (20b) right after (a).

(20) (a) John was winning the race.
(b) But then he fell and Mike won.

Sentences such as (21), which uses a stative verb with an adverbial that indicates a point expression, have been analyzed in different ways. Vendler (1957, p. 154) argues that the situation uses an achievement sense of "know"; more similar to how Moens and Steedman analyze this situation, Comrie (1976, p. 20) assumes an *ingressive* meaning, which points out the beginning of a situation.

(21) I suddenly knew.

This concludes our discussion of inherent lexical aspect, and we now turn to grammatical aspect or, as it is sometimes called, viewpoint aspect.

2.2 Grammatical aspect / viewpoint

We here review linguistic phenomena that have been treated as *grammatical aspect* or *viewpoint* (Smith, 1997). This aspectual opposition comprises categories such as *perfective* vs. *imperfective* and is stated independently of eventuality type. Perfective viewpoint presents situations as complete with both initial and final endpoint, while imperfective viewpoint makes only certain parts of the situation visible to the receiver (Smith, 1997, chap. 4). It is the perfective-imperfective distinction that has traditionally been referred to as *aspect* in Slavic linguistics (Filip, 1999). Especially the different subtypes of imperfectivity (Comrie, 1976), with habituals as a special class, are highly relevant for identifying some of Smith's situation entity types (Generic Sentence and Generalizing Sentence). Figure 2.2 shows Comrie's (1976) classification of aspectual oppositions that can be expressed by the verbal forms that refer to situations.

2.2.1 Perfective vs. imperfective

The most fundamental distinction in the hierarchy shown in Figure 2.2 is that between *perfective* and *imperfective* meaning. The latter is defined as making explicit reference to the internal temporal constituency of the situation, from 'inside'; perfective aspect presents a situation as a whole, from 'outside', i.e., without making reference to internal phases. In example (22), the situation referred to by the first clause is presented imperfectively, focusing on the middle phase of John's eating. The situation introduced by clause (b), in contrast, is viewed in the perfective. Here, the interpretation is that (b) happens during the time interval at which (a) is true.

(22) (a) John was eating a sandwich
(b) when Susan entered.

Smith (1997, p. 78) additionally assumes a *neutral* viewpoint, which is flexible in its interpretation. It includes the initial endpoint of a situation and at least one internal stage; and forms in neutral viewpoint allow both open and closed readings. Such a viewpoint occurs for instance with the French Futur. According to French native speakers, (23) can be interpreted in two ways (given in the English translation). The first translation is the perfective reading with an ingressive meaning (closed reading), the second reading is imperfective (open reading).

(23) John chantera quand Marie entrera dans le bureau.
John will start to sing / will be singing when Marie will enter the office.

The meaning of the English Perfect extends over mere temporal implications; it is thus an aspect in a sense different from the other aspects explained here (Comrie, 1976, chap. 3,

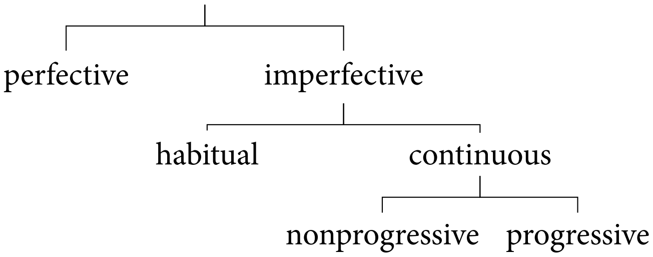

Figure 2.2: Classification of aspectual oppositions according to Comrie (1976).

Smith 1997, sec. 5.3.2). For example, in the **perfect of result** (24a), a present state ("John is in the US") is referred to as being the result of a past situation, and the **experiental perfect** (24b) states that a situation has held at least once during some time in the past.

(24) (a) John has gone to the US. (*perfect of result*)
(b) John has been to the US. (*experiental perfect*)

The classification in Figure 2.2 contains the most typical subdivisions of imperfectivity; many languages express imperfectivity using a single category or have categories that correspond to only parts of the meaning of imperfectivity (Comrie, 1976, p. 25). We next turn to the discussions of the subdivisions of imperfectivity.

2.2.2 Habituality

Habituals are sentences that "express regularities about the world which constitute generalizations over events and activities" (Carlson, 2005); on a sentence-level, they can be regarded as "derived statives" (Smith, 1997, p. 33). They have the interesting property that they allow exceptions, e.g., sentence (25) is still true if Mary eats something else once.

(25) Mary eats oatmeal for breakfast.

Habitual sentences "describe a situation which is characteristic of an extended period of time" (Comrie, 1976). It remains to decide what constitutes a characteristic feature rather than an accidental situation. This decision is of conceptual, not of linguistic nature. Habituality is sometimes confused with *iterativity*, which states that a situation occurred repeatedly. Examples (26) and (27) illustrate the difference. While there is a repeated situation in the first example, it is not habitual. Iteratives describe punctual events taking place several times in succession and are especially common for semelfactive verbs such as *cough* or *blink* (Smith, 1997). Such sentences are episodic in nature. Example (27), in contrast, uses the habitual form "used to" without there being any iterativity.

(26) The lecturer coughed five times.

(27) Simon used to believe in ghosts.

Habitual sentences may also use stative predicates (28), generalizing over situations in which some state applies (Smith, 2005, p. 5).

(28) Sloths sometimes rest on trees.

Habituals are not restricted to what one would usually consider a matter of habit (Carlson, 2005); they can also have inanimate subjects as illustrated by (29).

(29) Glass breaks easily.

It is worth noting that habitual aspect can be combined with other aspects. For example, in

English, the habitual "used to" construction can be combined freely with the Progressive, as illustrated by example (30) (Comrie, 1976, Smith 1997, p. 51).

(30) John used to be playing the viola (whenever I visited him).

We now turn to describing approaches to capturing the semantics of habitual sentences. Krifka et al. (1995, p. 30, 32) use the quantifier **Gen** for characterizing sentences that quantify over situations. Formally, they define habitual sentences as expressing generalizations over situations that are specified by the corresponding episodic verbal predicate.

(31) A sentence is *habitual* iff its semantic represention is of the form
Gen[...s...;...](**Restrictor**[...s...]; **Matrix**[**...s...**])
where s is a situation variable.

The representation of example (32) says that if there is a situation in which Mary comes home, she will smoke in that situation:

(32) Mary smokes when she comes home.
Gen[s;x](x=**Mary** & x **comes home** in s; x **smokes** in s)

In a different approach, Boneh and Doron (2013) represent habituality by means of the operator **Hab**, a modalized existential quantifier over sums of events. This operator shifts the verb itself to a habitual reading and stativizes the VP. In example (33), **Gen** generalizes only over individuals, the generalization over events is applied to the verb itself using **Hab**.[5]

(33) Women smoke.
Gen[x](x **is a woman**) (Hab e **smoke**(e, x))

Hab applies directly to the verb and can be input to both imperfective and perfective aspect. This is in contrast with other approaches (such as Comrie, 1976) where habituality is a realization of imperfective aspect. Boneh and Doron illustrate the difference of the English forms expressing habituality in the past, simple past, *used to* and *would* with regard to their perfective or imperfective interpretation using the following example.

(34) (a) In the eighties, John went to work by bus.
(b) In the eighties, John used to / would go to work by bus.

Sentence (34a) can be interpreted in a perfective way where John's habit is included in the reference time *in the eighties*. (34b) only has the interpretation of John's habit covering the whole eighties, hence it has imperfective aspect. Filip and Carlson (1997) also argue for the existence of perfective habituals. More generally, they claim that *sentential genericity*, which corresponds to habituality, is independent from tense and aspect. They observe that in languages with overt tense and aspect marking (such as Czech or Russian), both perfective and imperfective verb forms can be used to express generics.

[5] Example from Boneh and Doron (2013, p. 178).

Setting aside the question of whether habituals are part of the aspectual system or not, habituals generalize over events and require these to occur at least somewhat frequently, e.g., (35a) is only true if John is an active tennis player. In contrast, sentences such as (35b), which only denote abilities or preference but not actual events, are called *dispositional.*

(35) (a) John plays tennis. (*habitual*)
(b) John can play tennis. (*dispositional*)

More details on the interaction of negation and modal verbs with habituals will be explained in Chapter 5.

2.2.3 Progressive vs. nonprogressive

For the sake of completeness of our survey of aspectual oppositions, we briefly introduce the concept of *continuousness*, which means that something is not interrupted. Continuousness can be defined as imperfectivity that is not habituality (Comrie, 1976, p. 33), and has the two the two subcategories *progressive* and *nonprogressive* (see Figure 2.2). Progressiveness can be regarded as the combination of continuousness with nonstativity (Comrie, 1976, p. 12); thus, progressive constructions require a dynamic verb and are ungrammatical with stative verbs, see (36).

(36) He was running.
*He was knowing the answer.

Example (37) illustrates two nonprogressive cases: (a) is not continuous, i.e., not ongoing, and (b) is continuous but stative.

(37) (a) He ran.
(b) He knew the answer.

The English Progressive has extended its meaning well beyond the original definition of a combination of continuous meaning and nonstativity (Comrie, 1976, p. 38). For example, lexically stative verbs cannot be used with the Progressive. However, if they are used with the progressive form, they usually have a different meaning, e.g., in (38), the verb refers to a developing process (Comrie, 1976, p. 37). Lexically stative verbs can also be used in the Progressive if they refer to a temporary state (39).

(38) I'm understanding more about quantum mechanics as every day goes by.

(39) (a) I live at 6 Railway Cuttings. (*permanent state*)
(b) I'm living at 6 Railway Cuttings. (*temporary state*)

2.2.4 Marking of aspect

Aspectual distinctions are of semantic nature, and they may or may not be grammaticalized or lexicalized in individual languages (Comrie, 1976; Filip and Carlson, 1997). In linguistics, *markedness* refers to the following: if in an opposition with two or more members one is felt to be more usual or less specific, it is called *unmarked* (Comrie, 1976, p. 111). There are also categories in which all members may be equally marked. In some cases, the marked category signals the presence of some feature, the unmarked category says nothing about it. Sentence (40b) is such an unmarked case: it is not explicitly marked for habituality, but it does not exclude habitual meaning (Comrie, 1976, p. 124).

(40) John cycled to work.

For example, with a context indicating habituality, we can replace the English simple past with the habitual form "used to," as shown in (41).

(41) John cycled / used to cycle to work for five years.

Carlson (1995) suggests that habitual sentences are formally based on episodics, as there is a corresponding episodic ("The sun rose in the East") for any generic ("The sun rises in the East"). Based on a study of the syntactic marking of the episodic-habitual distinction in tense-aspect systems of 65 languages, Dahl (1995) remarks that this misleadingly suggests that the generic interpretation is the syntactically marked case. In fact, according to her study, generics have a tendency to be minimally marked for tense and aspect. As an example, English generics use the Simple Present, while episodics are marked with the Progressive. Possible explanations for this *minimal marking tendency* is that generics usually do not have a reference in time, or that generics may occupy a region of the semantic space that is not close to any of the origins of morphemes and grammatical constructions marking tense and aspect (Dahl, 1995, p. 416). This means that for automatically identifying generics, a variety of factors, grammatical and lexical, needs to be taken into account. Dahl also notices that there are exceptions to the minimal marking tendency, such as Hindi.

Dahl also remarks upon two different types of habitual contexts. In all languages of her study, overt marking of habituals is either obligatory in (i) and optional in (ii) or optional in (i) and absent in (ii).

(i) Cases where *usually* can be inserted: "What does your brother (usually) do after breakfast?"

(ii) Cases where *usually* can not be inserted: "What kind of work does he do?"

For marking aspectual oppositions, some languages use morphological (synthetic) means, e.g., Chinese has the morphological marker "-zhe" for the Progressive. In contrast, the English Progressive uses syntactic (analytic) means – copular verb + predicate – to formally express aspectual distinctions (Comrie, 1976, pp. 87–88).

However, there is no one-to-one correspondence between the English Progressive and imperfective meaning; in English the opposition of perfective and imperfective meaning has not been grammaticalized. The Progressive/Nonprogressive distinction is only comparable to the imperfective/perfective distinction for nonstative verbs and then only if excluding habitual meaning (Comrie, 1976, p. 7). There are many language-particular categories that correspond closely but not exactly to semantic distinctions. The English Progressive usually expresses progressive meaning, but as explained above, its use is somewhat wider (Comrie, 1976, p. 10). The Simple Present tense in English usually invokes a habitual reading. It may also be used in particular narrative modes where usually the Progressive would be used (Comrie, 1976, p. 68, 73, 77; Smith, 1997, p. 111). In some languages, aspectual distinctions are restricted to one or more tenses, i.e., they do not operate independently of tense (Comrie, 1976, p. 71).

2.3 Genericity

There are two distinct phenomena which have been referred to as *genericity* in the linguistic and semantic theory literature (Krifka et al., 1995). One of them is the generalization over episodes, and we have introduced this phenomenon above in Section 2.2.2 as habituals. The second phenomena is *reference to a kind* as in (42).

(42) Dinosaurs were huge.

Both phenomena are relevant for our corpus-linguistic work on situation entity types; they also often co-occur. Reference to a kind is a feature of noun phrases (NPs) while habituality is a feature of the entire clause. This section gives an overview of genericity, for the most part and if not otherwise stated following Krifka et al. (1995).

2.3.1 Reference to kinds

Krifka et al. (1995, p. 14) introduce the term *kind-referring* for noun phrases that refer to kinds (43a) rather than objects (43b).

(43) (a) The lion is a carnivore. (*kind-referring*)
(b) The lion escaped from the zoo yesterday. (*object-referring*)

Kinds are assumed to be a certain type of individual entities which can be referred to (Krifka et al., 1995, p. 65). Particular individuals can belong to a kind, e.g., "Simba" can be a particular individual belonging to the kind "lion". What exactly constitutes a kind and what an object depends on the cultural context. Krifka et al. (1995) suggest that there are *well-established kinds* such as "the Coke bottle", giving "the green bottle" as a counterexample (Carlson, 1977a).

Krifka et al. assume that in English, only definite singular count nouns, bare plural count nouns and bare mass nouns can be considered as kind-referring. However, it has been argued that indefinite singular nouns can also be generic. Lawler (1973) distinguishes *definite generics*, which are expressed by definite singular noun phrases as in example (6), from *indefinite generics*. The latter are expressed by indefinite singular noun phrases as in (7).

(44) The university is no place to fight a war. (*definite generic*)

(45) A telephone can be either a help or a nuisance. (*indefinite generic*)

According to Krifka and Gerstner (1987), in the case of definite generics the statement applies to the kind, to say for instance "A telephone was invented" is to make a category mistake. It is, of course, grammatical if assuming a taxonomic reading, i.e., "a telephone" refers to a "type of telephone". Sentence (46) shows an example for such a *taxonomic reading* (Krifka et al., 1995, p. 5).

(46) The World Wildlife Organization decided to protect a (certain) large cat, namely the Siberian tiger. (*kind-referring*)

In the case of indefinite generics as in (45), on the other hand, the property attributed to the kind could in principle also be held by individuals of that kind (see also Asher and Morreau, 1995, p. 300).

Lawler (1973) notices that if definite generics are in subject position, the verb phrase must be either stative or a *kind predicate* ("invent", "be extinct" etc.), as otherwise a generic interpretation is not possible. Compare (47a) and (47b).

(47) (a) The oppossum hangs by its tail. (*definite generic*)

(b) The oppossum hangs by its tail this afternoon. (*object-referring reading*)

As can also be seen by (47), genericity as reference to kinds is not in the NP itself: it depends on the clause how the NP is interpreted.

2.3.2 A cross-classification of generic phenomena

Krifka et al. (1995, p. 14) propose a cross-classification of generic phenomena as shown in Table 2.3. We present and explain this classification here as it shows the breadth of cases related to genericity, thus illustrating distinctions that will be important for the development of our situation entity type annotation scheme.

A. Habitual sentences

(i) Simba (usually) roars when he smells food. (*specific non-kind-referring*)
(ii) The lion (usually) roars when it smells food. (*specific kind-referring*)
(iii) A lion (usually) roars when it smells food. (*nonspecific non-kind-referring*)
(iv) A predatory cat (e.g. the leopard) (usually) is exterminated when it is dangerous to people. (*nonspecific kind-referring*, taxonomic reading)

B. Lexical characterizing sentences

(i) Simba has a mane. (*specific non-kind-referring*)
(ii) The lion weighs more than most animals. (*specific kind-referring*)
(iii) A lion (usually) weighs more than 200 lbs. (*specific kind-referring*)
(iv) A predatory cat (e.g., the lion) (usually) knows its young. (*nonspecific kind-referring*, taxonomic reading)

C. Episodic dynamic sentences

(i) A lion attacked a visitor yesterday. (*nonspecific non-kind-referring*)
(ii) Simba roared. (*specific non-kind-referring*)
(iii) The lion disappeared from Asia. (*specific kind-referring*)
(iv) A predatory cat (e.g., the Siberian tiger) disappeared from Asia. (*nonspecific kind-referring*)

D. Episodic statives

(i) Simba is in the cage. (*specific non-kind-referring*)
(ii) A lion is in the cage. (*nonspecific non-kind-referring*)

Figure 2.3: A cross-classification of generic phenomena (Krifka et al., 1995, p. 34).

The first dimension across which sentences are classified is whether they are habitual or episodic; the latter comprises both episodic dynamic sentences in which something happens and episodic statives which describe temporary states. There are also stative sentences that, due to the lexical meaning of their predicates, *characterize* their subject, which may or may not refer to a kind (48).

(48) (a) Giraffes are tall. (*kind-referring*)
(b) Bob is tall. (*non-kind-referring*)

Krifka et al. also introduce the distinction of *specific* vs. *nonspecific*, where the former includes all cases in which an NP refers to a particular individual. This distinction is independent of the kind reference vs. object reference distinction. The subject NP in (49a) does not directly refer to a kind but to an arbitrary member thereof; in our annotation scheme, we thus treat such cases as generic.

(49) (a) A lion has a bushy tail. (*nonspecific*, *non-kind-referring*)
(b) The lion is a carnivore. (*specific*, *kind-referring*)
(c) Simba is a lion. (*specific*, *non-kind-referring*)
(d) A lion must be standing in the bush over there. (*specific*, *non-kind-referring*)

Kind-referring NPs may occur in episodic dynamic sentences (Wilkinson, 1995, p. 386) as in (50). In this case, it follows that something is true for the kind because it is true for some specific members of the kind.

(50) (a) I saw bears in the zoo. (*specific non-kind-referring*)
(b) I saw that kind of animal in the zoo. (*specific kind-referring*)

2.3.3 Semantic interpretation of generics

In this section, we briefly review some approaches to capturing the semantics of generic sentences. We have already given Krifka et al.'s representation for habitual sentences in Section 2.2.2. They distinguish characterizing (including habitual or kind-referring) sentences from particular sentences as the former must have at least one variable to generalize over (Krifka et al., 1995, p.32). In (51), the operator **Gen** generalizes over the members of a kind.

(51) Unicorns have horns.
Gen[x;y](x **are unicorns**; y **are horns** & x **has** y)

Generic sentences express regularities within classes of entities, and thus are similar to universally quantified sentences in their truth conditions and entailment properties. However, their truth-conditional interpretation is tricky, since they express typicality, describe stereotypes and allow exceptions, for example "Dutchmen are good sailors" is not false even if most Dutchmen do not sail at all (Carlson, 1977a). Carlson (1995) identifies two approaches to interpreting generic sentences. In the *inductive view*, generics are inductive generalizations, one needs to observe many instances of episodes in order to make the generic true, e.g., "dogs bark." The *rules-and-regulations (realist) view* assumes that generics are structures that are not episodic instances but it is causal forces that make the statement generic. Evidence for this view is that some things are simply defined, for instance the rules of a game ("Bishops move diagonally").

2.4 Role of aspect in theories of discourse

Smith (2003) developed the set of situation entity types which are central to this thesis in the context of her work on *discourse modes*. Situation entities are discourse entities that are aspectual in nature. To give a broader picture, in this final part of our chapter reviewing the related linguistic theory, we address the relationship between aspect and discourse as it has been described in linguistics.

2.4.1 Discourse modes and text types

Smith's main motivation for introducing her inventory of situation entity types is a text-typological one. Without claiming exhaustiveness, Smith (2003) introduces the five *modes of discourse* **Narrative**, **Description**, **Report**, **Information** and **Argument/Commentary** with respect to two features (listed in Table 2.3): (a) their principles of progression, which may be temporal or atemporal, and (b) their usage patterns of the situation entity types. Each of these features, in turn, has linguistic correlates.

Discourse modes	**progression**	**predominant situation entity types**
Narrative	temporal	State, Event
Report	temporal	State, Event, General Stative
Description	spatial	State, Event
Information	metaphorical	General Stative
Argument	metaphorical	Fact, Proposition, General Stative

Table 2.3: Linguistic features of discourse modes (Smith, 2003).

In the Narrative mode, progression corresponds to advances in narrative time; in the Report mode, which is also temporal, advancements are always anchored to the speech time. A spatial advancement through a scene or object indicates Description mode. Metaphorical progression in the Information and Argument mode does not mean that metaphors in the usual sense have to be used; this rather means that progression is similar to the spatial progression of the Description mode, i.e., it advances through the domain of the text.

In addition to Smith's modes of discourse, there are at least two related theories of *text types* that were developed independently for German and French. Like the discourse modes, these theories capture text types as a categorization explicitly orthogonal to *genre*, the classification of text types according to similar form or content (Freedman and Medway, 1994).

Werlich's (1989) *text types* understand texts as sequences of semantically coherent sentences. Texts usually open with a *Sequenzinitiator* (sequence initiator), then have several *Sequenzsignale* (sequence signals) and end with a *Sequenzterminator* (sequence terminator). The *dominant sequences* of the various text types are the result of developing the obligatory initiators which differ per type. For example, Werlich's *descriptive* text type opens with a "phenomenon-registering sentence," which corresponds to a generic sentence. His argumentative text type starts with relations between concepts or sentences that attribute a quality to something. Similar to Smith's discourse modes, the sequence forms of his text types follow some principle of progression, e.g., temporal structure for narrative texts or locality information for descriptive texts.

Adam (2011) also proposes five *type de texts* (text types): *narratif*, *descriptif*, *explicatif*, *argumentatif* and *dialogal*. His text types have internal structure, and incorporate more features from traditional genre studies than Smith's and Werlich's work; e.g., his argumentative mode consists of a Thesis, Argumentative phase and a Conclusive phase. Linguistic features such as reported speech or presence of modals are given for the various text types.

2.4.2 Temporal interpretation

As we have already shown in Chapter 1, clause-level aspect and discourse-level semantics interact such that the reader or hearer may arrive at a temporal interpretation of the events and states mentioned in the discourse. According to Dowty (1986), in the absence of definite time adverbials or other pragmatic factors that override default interpretation, the aspectual classes of the predicates in a discourse determine the temporal relationships between the events and states they describe. Dowty takes aspectual class as an attribute of sentences rather than lexical items, but states that his analysis would also work if events or situations were taken as the primitives. In his account, the determination of aspectual class relies on semantic properties of verbs rather than syntactic properties as in earlier work (e.g., Vendler, 1957).

The default interpretation for sentence pairs where the first sentence describes an accomplishment or an achievement is that the second sentence is interpreted as describing an event that occurs after the event described in the previous sentence, as Dowty illustrates using example (52).

(52) John entered the president's office. The president walked over to him.

If the second sentence has a stative predicate, an activity predicate as in (53) or a progressive construction, however, narrative time does not move in the second sentence, and the respective state or process is interpreted as overlapping with the events described by the surrounding discourse.

(53) John entered the president's office. The president was writing a letter.

Dowty stresses that any interpretation of duration of events requires a lot of common sense, and that a theory of temporal interpretation of sentences in a discourse must take into account pragmatic principles. From this follows that the difference between accomplishments and achievements is almost negligible; what matters for his account is that both have natural endpoints. Accomplishments are usually described as having some duration, while achievements are "punctual." However, the latter often also have some duration, depending on the hearer's interpretation. Whether a predicate is classified as an accomplishment or achievement, hence, depends on the granularity of the assumed ontology of event structure.

It is important to remember that the distinction between closed and open (or perfective and imperfective) situations pertains to narrative time rather than real time (Smith, 1997, p.66). The presence of endpoints has consequences to the temporal interpretation in discourse; when trying to place the constituent introduced by the next sentence within the narrative time structure that we have already constructed, we are able to use end points (Smith and Erbaugh, 2005; Kamp and Rohrer, 1989, as cited by Smith (1997), p. 66).

Chapter 3

Related work in computational linguistics

This chapter provides a survey of the computational work related to the experiments and corpus presented in this thesis. We first review approaches to automatically classifying the aspectual class of verbs or situation-denoting phrases or clauses (Section 3.1). The second part of this survey addresses computational approaches to identifying generic expressions. We explain the respective annotation schemes and also comment on agreement and problematic cases in Section 3.2.2. Closely related previous work includes methods for recognizing habituals (Section 3.2.1) and for NP-level genericity (Section 3.2.3).

3.1 Automatic classification of aspectual class

In this section, we survey the work in computational linguistics that addresses the computational modeling of aspectual class in various ways. Early studies (Nakhimovsky, 1988; Passonneau, 1988; Brent, 1991; Klavans and Chodorow, 1992, see Section 3.1.1) laid foundations for a cluster of papers published in the late 1990s (Siegel, 1998b,a; Siegel and McKeown, 2000). Since then, it has mostly been treated as a subtask within temporal reasoning, such as in efforts related to TimeBank (Pustejovsky et al., 2003b) and the TempEval challenges (Verhagen et al., 2007, 2010; UzZaman et al., 2013). The work on automatic aspectual classification presented in this thesis is most closely related to the work by Siegel and McKeown (Section 3.1.2) and the work by Palmer et al. (Section 3.1.4), who also explicitly model Smith's (2003) situation entity types.

3.1.1 Early studies

Coming from an Artificial Intelligence (AI) perspective, Nakhimovsky (1988) maps out the structure of knowledge sources that would be required for an implementation of narrative understanding. He includes knowledge about the internal constituency of events and the temporal relations between them along with knowledge about usual durations.

He defines *aspect* as a grammatical category of the verb and *aspectual class* as a characteristic of lexical meaning, i.e., situation or event types. His lexical entries for aspectual class are called *h-types* (for *histories*) and consist of a preparatory stage, an initial stage (or initial point), the body, the final stage (or final point) and a resulting stage. Finally, the *aspectual perspective of the sentence* is determined by the position of the reference time (RT) with respect to the phases of the h-type. For example, "they were eating strawberries" puts the RT inside the body, while "they had eaten strawberries" focuses on the resulting phase. Nakhimovsky's work is related to that of Moens and Steedman (1988) explained in Section 2.1.4, and appeared in parallel. Similar to Moens and Steedman's inventory of aspectual types, Nakhimovsky's h-types comprise instantaneous events, states, atelic processes and telic processes.

In the PUNDIT system for temporal information processing (Passonneau, 1988), the temporal structure of tensed clauses is represented as one of the situation types *state*, *event*, *process* or *transition event*. The system parses each sentence, extracting, for each verb, the following features: tense, presence of Perfect or Progressive and arguments, as well as the decomposition (a lexical entry) of the verb. From this decomposition, lexical aspect can be read off: lexical entries of transition events use `become`, those of processes use `do`; other cases signal states. Using lexical aspect and the information of whether the verb is in the Progressive, situation type is determined as input to the module computing within-sentence temporal location, which are represented in a manner similar to Reichenbach's (1980) analysis of tense.

Klavans and Chodorow (1992), in the context of lexicon induction, suggest representing the event structure of a verb as its *degree of stativity*. A verb type's degree of stativity is estimated by the proportion of occurrences of the verb type in a corpus that are in the Progressive, based on the assumption that stative verbs are less likely to occur in the Progressive. They compare the values of the 100 most frequent verbs in the Brown corpus and those of the 115 most frequent verbs in the automatically parsed 1-million words Reader's Digest corpus to the literature on stative verbs, thereby confirming their intuitions. Brent (1991) also presents a program for identifying stative verbs using two syntactic indicators: the Progressive and whether verbs combine with rate adverbs such as "quickly" and "slowly." His evaluation consists of the inspection of the system's output for the 204 verb types occurring at least 100 times each in the one-million words Lancaster/Oslo/Bergen (LOB) corpus, and he finds promising results.

Dorr (1997) develops a database of about 4000 English verbs, categorized according to Levin's (1993) verb classes. More details on Levin's verb classes will be given in Section 3.1.5. Dorr and Olsen (1997) show that from the features given in the LISP-structured lexical entries for the verbs, it is possible to read off telicity, dynamicity, and durativity. For example, dynamicity is characterized by the entry at the topmost level of the lexical conceptual structure for a verb: entries for events use `go`, `act`, `stay`, `cause` or `let`, while entries for states use `go-ext` or `be`.[1] We make use of this strategy in order to derive seed sets for classifying lexical aspectual class of verbs in the experiments explained

[1] The structural primitive `go-ext` is used for verbs describing extensions, e.g., "The road extended from NY to CA.", see also `http://www.umiacs.umd.edu/~bonnie/Demos/LCS_Database_Documentation.html`

in Section 8.2.

3.1.2 Linguistic indicators for aspectual classification

Siegel and McKeown take the above ideas one step further by leveraging various linguistic markers of aspect rather than just the Progressive, following ideas of Dowty (1986). They automatically classify the aspectual class of verbs, including stativity and telicity/completedness, and introduce the use of linguistically based numerical indicators (Siegel, 1998b; Siegel and McKeown, 2000). These linguistic indicators can be learned from automatically parsed corpora using aspectual markers. The English Progressive, which is more likely to occur with events rather than with states, is one example for such a marker. For each verb type, normalized counts are obtained for the various linguistic markers (see Section 7.1.6 for the complete set) from the background corpus. The aim of this work is to classify the *fundamental* aspectual class of a verb in context, which is a function of the verb and a select group of arguments and modifiers, which may differ per verb. The fundamental aspectual class of a verb may differ from the clause's aspectual class, as illustrated by the following examples (taken from Siegel, 1998b, sections 2.1.6-7):

(1) (a) I stared at it. (**non-culminated process** → *atelic*)
(b) I stared at it <u>for 10 minutes</u>. (**culminated process** → *telic*)

The prepositional phrase (PP) "for 10 minutes" indicates the duration of the non-culminated process of "staring," coercing the clause to include the endpoint (see also Moens and Steedman, 1988). The fundamental aspectual class of (1b) is still that of a non-culminated process. According to Siegel (1998b), a natural language understanding system must first recognize a clause's fundamental aspectual category and can then determine which aspectual transformations have affected the clause. To give another example, correct identification of a verb's fundamental aspectual class is a prerequisite for interpreting the meaning of *for*-PPs. With a non-culminated process, the *for*-PP indicates the duration of the process (2a); with a culminated event, it denotes the duration of the resulting state (Siegel and McKeown, 2000, section 2.3).

(2) (a) I stared at it <u>for an hour</u>. (**non-culminated process**)
(b) I left the room <u>for an hour</u>. (**culminated event**)

Siegel and McKeown use small labeled data sets (308 instances for stativity and 739 instances for telicity) for training in a machine learning step that learns which combinations of linguistic indicator values should classify a verb as stative or telic. Both data sets are labeled in a binary way as **yes/no** for the two classes. The labeled test sets for both classification tasks have the same sizes as the respective training sets. The data labeled for stativity is taken from medical discharge summaries; the data labeled for completedness is taken from a set of ten novels and excludes any stative clauses. Clauses whose main verb is "be" or "have" are excluded because the former are always stative and the latter are highly ambiguous (Siegel, 1998a). Linguistic indicator values are computed over the respective data sets in each case.

Siegel and McKeown compare three machine learning methods, namely genetic programming, logistic regression and decision trees, of which the latter method is found to work best. In the case of stativity, the system achieves an accuracy of 93.9%, performing better than a baseline using the overall majority class in the data set (83.8%) but worse than a system memorizing the most frequent class per verb type (94.5%). For telicity, the overall majority class (**yes**) is 63.3%, while a baseline using the most frequent class per verb type results in an accuracy of 70.8% and the system using linguistic indicators achieves 74.0%.

For verb types for which training data exists, this approach never performs better than using the majority class for each verb type. Siegel and McKeown (2000) argue that aspectually categorizing verbs is a first step towards aspectually classifying clauses, and that the most frequent category of a verb simply needs to be determined for each domain. In addition, it is possible to classify instances of verb types that occur in the test set, but not in the training set. This is the reason why the system is able to outperform the baseline memorizing each verb type's most frequent class in the case of completedness, but not for stativity: half of the instances in the completedness test set have verb types for which no training data exists, but only about 15% of the test instances labeled for stativity do not have training data for the respective verb types.[2]

Siegel (1998a) take first steps in automatically determining the aspectual class of aspectually ambiguous verb types using the WordNet (Fellbaum, 1998) category of the verb's direct object. They mark a total of 206 clauses from the corpus of medical discharge summaries with the main verb "have" according to stativity and divide them equally into a development set and a test set. The nouns occurring as the direct objects are each placed into one of the 25 categories at the top of WordNet's semantic hierarchy, and the additional category `pronoun` is used for pronouns. Based on the development set, a rule is developed manually which labels clauses as events if their direct objects belongs to the categories `event`, `act`, `phenomenon`, `communication`, `possession` or `food` and as states otherwise. On the test set of 103 clauses, this method achieves an accuracy of 79.6% compared to a majority class baseline of 69.9%. The method cannot perform better than an upper bound of 84.5% because in the remaining cases, the WordNet category of the direct object occurs in both stative and eventive test cases, i.e., one of them will be predicted wrongly. As an additional problem, Siegel (1998a) identifies word sense disambiguation: in the work described above, they simply used the most frequent sense for each noun. Siegel (1998a) also suggest computing linguistic indicators for combinations of verb types and the WordNet category of the direct object instead of for verb types only. Preliminary experiments on the completedness distinction and data described above are promising: while accuracy is comparable, the precision-recall trade-off seems favorable when using the object categories in addition.

The above work by Siegel and McKeown is based on relatively small data sets, but has been a major inspiration for the work presented in this thesis (see Section 8.2; Friedrich and Palmer, 2014a) as well as other works. The ideas have been adapted to Chinese (Cao et al., 2006), and Hermes et al. (2015) use similar ideas to induce Vendler classes at the

[2] Computed from the data available at `http://www.cs.columbia.edu/~evs/VerbData`, which contains the verbs, linguistic indicators and labels of each instance. Unfortunately, the original text data for each clause was not preserved.

type level for 95 German verbs.

3.1.3 Event classes in TimeBank

Interest in temporal and event-based reasoning rose again in the early 2000s. In this context, the TimeBank corpus (Pustejovsky et al., 2003b) was manually annotated with events, times and relations that hold between them according to the TimeML guidelines (Pustejovsky et al., 2003a). TimeML events are "situations that occur or happen," but include also "states or circumstances in which something obtains or holds true." Thus, the usage of the term *event* here is more similar to Smith's term *situation entity* than to her Event.

TimeML events can be expressed by tensed verbs, stative adjectives and event nominals. Only situations that are temporally located in the text are marked and linked, thus excluding generics. Each event is annotated with its *event class*, which can take one of the following values (Saurí et al., 2005b), which are assumed to be helpful for determining factuality of the event (Saurí et al., 2005a). Correspondences between TimeML event class and situation entity types are noted where relevant.

- OCCURRENCEs happen, corresponding to the situation entity type Event (*die, crash, merge, sell*).
- STATE describes circumstances in which something holds (*like, own*); they are broader than Smith's State as they are also annotated for prepositions (*on board*) or adjectives (*the kidnapped girl*).
- I_ACTION is the label for intensional actions (*try, persuade, swear*).
- I_STATE is the label for intensional states (*love, believe, enjoy*).
- ASPECTUAL predicates pick out a phase of the EVENT that they take as their argument (*begin, start, continue*).
- REPORTING is for capturing attribution, corresponding to the situation entity type Report (*say*).
- PERCEPTION events involve the physical perception of another event (*see, hear, feel*).

The TempEval challenges (Verhagen et al., 2007, 2010; UzZaman et al., 2013) are a series of shared tasks aiming at the "automatic identification of temporal referring expressions, events, and temporal relations within a text." The top-performing systems (Jung and Stent, 2013; Bethard, 2013; Chambers, 2013) use corpus-based features, WordNet synsets, parse paths and features from typed dependencies to classify events as a joint task with determining the event's span. There is also work automatically recognizing which of the above event classes an event belongs to. Like situation entity type classification, this requires a combination of various lexical factors. In contrast to situation entity types, however, this task operates completely at the word sense level. Saurí et al. (2005a), in their event recognition system, simply assign the class that was most frequently observed for each verb type in the training data to events and reach an accuracy of 82.3% on TimeBank 1.2. Bethard and Martin (2006) phrase the recognition of EVENTs and their semantic class as a chunking task using the B-I-O formulation. Labels indicate that a word is *outside* (O) any event mention, the *beginning* (B) of or *inside* (I) an EVENT mention (e.g., B_OCCURENCE, I_OCCURRENCE). They use a wide range of syntactic-semantic features, including among others morphological features, root verbs, WordNet hypernyms, word cluster, determiner type, governing temporal prepositions or part-of-speech. Training on 90% of TimeBank and testing on the remaining 10%, when evaluating for verbs only, they reach a precision of 86.4% and a recall of 90.3%; for event *and* class identification, they reach 71.4% and 70.1% respectively. Their ablation tests show that affixes, word cluster information and WordNet features are most important. Llorens et al. (2010) extend this idea by using a conditional random field enhanced with semantic role information. Their F1-score for event classification is 64.3%, not directly comparable to the above systems as they use 5-fold cross validation on TimeBank. Derczynski and Gaizauskas (2015) add features based on Reichenbach's (1947) interpretation of English tenses to their system for automatically identifying temporal relations. For their experiments, they use the gold-standard tense and aspect annotations as present in TimeBank.

Costa and Branco (2012) explore the usefulness of a wider range of explicitly aspectual features for temporal relation classification. They obtain counts for linguistic indicators designed for Portuguese verbs from web queries, roughly following the ideas of Siegel and McKeown (2000). The features for the most common 4000 Portuguese verbs are then included as features in the various TempEval tasks of classifying temporal relations, leading to small but promising improvements.

3.1.4 Situation entity classification

The work presented in this thesis is most closely related to and builds on that of Palmer et al. (2004, 2007) on automatically classifying situation entity types in text.

In a first attempt, Palmer et al. (2004) distinguish Eventualities, Generalizing Statives and Abstract Entities, assigning labels to both nouns and verbs. Their data set consists of a gold standard of three texts from the National Geographic magazine, corresponding to approximately 200 annotated situation entities. Their system processes the texts using the XLE parsing system with the ParGram LFG grammar (Butt et al., 2002).[3] Linguistic tests check for the presence of features that are indicators for particular situation entity types, e.g., the presence of bare plurals is an indicator for generics. A set of manually ordered transfer rules checks for the presence of certain feature combinations in the parser's output and assigns a situation entity type based on them. In addition, lexical resources are used for deriving more features. Information on whether verbs are stative vs. dynamic and telic vs. atelic is taken from a database of Lexical Conceptual Structures (Dorr, 1991; Dorr and Olsen, 1997). In addition, Palmer et al. compile a list of factive and propositional predicates in order to facilitate the recognition of Abstract Entities. They find that including information from lexical resources improves recall but lowers precision for the situation entity classification task. Without lexical information, they reach a precision of 69.8% and a recall of 56.4%; when including lexical information, 65.4% and 59.6% are reached respectively.

[3] `http://www2.parc.com/isl/groups/nltt/xle`

Palmer et al. (2007) present the first data-driven model for the classification of situation entities using the inventory of Smith (2003): State, Event, Report as a subtype of Event, Generic Sentence, Generalizing Sentence and Abstract Entity (Fact and Proposition). They add Question and Imperative as Speech Mode types. Their model applies a ten-way classification, including None for clauses not invoking a situation such as headings or mentions of authors.

Palmer et al. distinguish between *basic situation type* and *derived situation type*. The *basic situation type* is determined by the verb and its arguments. The underlying basic situation type of (3) is the Event *Mickey paint house*, the use of the simple present results in an *aspectual coercion* to the derived situation type. In this thesis work, we use the situation entity type labels exclusively at the clause level; Palmer et al.'s basic situation types correspond to our level of analysis of lexical aspectual class, see Chapter 5.

(3) Mickey paints houses. (Generalizing Sentence)

A selection of texts from the popular lore section of the Brown corpus (Francis and Kučera, 1979) and from the Message Understanding Conferences 6 (MUC-6, Grishman and Sundheim, 1996) were annotated by two experts and adjudicated by a third. Segmentation was done manually for the Brown texts by one expert, resulting in 4390 clauses. The MUC-6 data were already segmented into elementary discourse units (EDUs), amounting to a total of 1675 clauses. About 10% were held out as test data. Automatic preprocessing was done using the C&C toolkit (Curran et al., 2007), providing part-of-speech tags and Combinatory Categorial Grammar (CCG, Steedman, 2000) categories for words and syntactic dependencies. Based on these, the following features are extracted:

- Words: words and punctuation in the clause.
- Words & Tags: words and punctuation, part-of-speech tags of each token, and word/part-of-speech tag pairs for each token.
- Linguistic Correlates: these features encode linguistic cues expected to be correlates of certain SE types (in the literature on SEs). For example, clauses embedded under the predicate *force* are usually Events.
- Grammatical relations: These features are extracted from the CCG parses of each clause, providing a deeper level of syntactic analysis such as identification of the main verb, the grammatical function of arguments and CCG categories.

Palmer et al. (2007) compare two models: (a) a maximum entropy model that simply labels each clause, and (b) a sequence labeling model that tags sequences of utterances, taking into account previously-predicted labels ("lookback features") and features of adjacent utterances. They adapt the OpenNLP maximum entropy part-of-speech tagger for this task. As additional features in their sequence model, they use the labels of up to six preceding clauses.

The most frequent label in the training set is State. When using this label as a simple majority class baseline, accuracies are 35.3% and 36.2% for Brown and MUC respectively. Using the Words features results in a accuracy of 45.4% on Brown, Words & Tags results in 49.9%. Adding the Linguistic Correlates does not result in an improvement. The authors hypothesize that more training data is needed to show an effect for such features. Using more deep syntactic information (CCG supertags and correct identification of the main verb rather than simply using the first verb in each clause as it is done in the other feature sets) raises accuracy to 50.6%.

When adding the gold standard labels of the previous clauses as features, accuracy improves steadily for each feature set. When using automatically predicted labels of previous clauses as features, accuracy improves when using one or two preceding labels, and then decreases. Training on Brown and testing on MUC-6 and vice versa, Palmer et al. find that sequence information only helps when training in-domain. They come to the preliminary conclusion that situation entity patterns are specific to the domain, genre or discourse mode. We reimplement and compare to the system of Palmer et al. in Section 8.5.1.

3.1.5 Other recent work modeling situation types

In this section, we give a brief overview of other recent computational approaches to modeling situation type, covering work that uses various situation type inventories.

Modeling of Vendler classes. Zarcone and Lenci (2008) build computational models of event type classification in context for Italian. They manually annotate 3129 occurrences of 28 Italian verbs with one of four event types corresponding to Vendler's verb classes *state, process, achievement* and *accomplishment*. 583 instances in their data set are stative. Three additional annotators mark 100 instances; accuracy versus the primary annotator ranges from 44% to 73%. They train a maximum entropy classifier using adverbial, morphological and syntactic features, as well as features capturing argument structure. They evaluate accuracy, precision and recall using 10-fold cross validation. They present results for a four-way classification task, and for two-way classification tasks for the features of telicity, durativity and dynamicity, dimensions along which the four event types can be classified. For dynamicity, which corresponds to our classification of stativity or lexical aspectual class, their classifier reaches an accuracy of 92% for the whole corpus, with a baseline of 88%.

In the context of the Richer Event Description (RED) annotation scheme (Ikuta et al., 2014), recently, the annotation of events with Vendler-style situation types has been proposed (Croft et al., 2016). The suggested annotation scheme includes subtypes for each of four Vendler categories, e.g., distinguishing between stage-level and individual-level predicates for states.

Modeling of Leech's classes. Keelan (2012) performs an automatic eight-way classification of verbs into the classes listed in Table 3.1, which are based on Leech's (1971) verb classes. While acknowledging that for classifying lexical aspect, the phrasal level, i.e., the verb's arguments and modifiers needs to be taken into account, Keelan conducts his classification task at the verb-type level, assuming the predominant lexically specified aspect for each verb, i.e., its most frequent sense. In the first set of experiments he

Keelan's verb class	Leech's verb class	Examples
Transitional Events	Transitional Event	hit, jump, nod, kick
Momentary Events	Momentary	arrive, die, fall, stop
Activity	Activity	drink, eat, play, rain, run
Change	Process	change, learn, develop
Perception	Inert Perception & Bodily Sensation	feel, hear, see, smell
Cognition	Inert Cognition	believe, forget, guess
Attitude	Attitude	hate, hope, like, prefer
Relationship	having and being verbs	be, belong, own, resemble

Table 3.1: Keelan's (2012, p. 4, 19) verb classes based on Leech's (1971) classes.

collects more prototypical instances of verbs for each of the verb classes by comparing the distributional contexts of verbs in Wikipedia to a small set of 65 seed verbs given by Leech. Manual analysis of the collected verbs by nine human judges reveals that the system chose the correct class for 79% of the verbs. However, human annotators achieved a relatively low agreement score of $\kappa = 0.29$ (Fleiss, 1971). After filtering according to majority support by the human judges, 155 seeds are retained as the basis of the second set of experiments, which aim to predict the correct class for each verb type in a 10-fold cross validation setting. The supervised classification setting uses a support vector machine (SVM, Cortes and Vapnik, 1995). The most important features were those describing the verb itself, i.e., counts in which aspects and tense the verb appears, and those describing prepositional phrases occurring in the contexts of the verb. Features describing adverbial phrases and co-occurring nominal arguments did not have an effect. The most frequent class baseline (Activity) reaches a macro-average F1-score of 31%. The system reaches a macro-average F1-score of 48% with an accuracy of 60%, the F1-score for the best-performing class (Transition) being 71% and the F1-score for the worst-performing class (Attitude verbs) amounting to 19%.

Modeling of event types for tense prediction. Correct choice of modals, tense and aspect remains a difficult problem in machine translation. As Chinese does not have tense, when translating from Chinese to English, it is a hard task for machine translation systems to pick the correct tensed form. With the aim of automatically inferring the "semantic" tense of events for Chinese, Xue and Zhang annotate tense, event type and modality on Chinese text (Zhang and Xue, 2014; Xue and Zhang, 2014). Xue et al. (2008) observe that annotating semantic tense – as past, present, future, relative past, relative present and relative future – on Chinese text directly is hard. Xue and Zhang (2014) hence employ a distant annotation approach, marking up the semantic tense, event type and modality of Chinese events on the English side of a word-aligned parallel English-Chinese corpus (Xue et al., 2005). The assumption is that more consistent annotations will be obtained this way as Chinese has no grammatical tense and English has richer morphosyntactic indicators for the labeled categories. In addition to tense, each span is labeled with one of the labels shown in Table 3.2 for *event type*. Habitual Events are events that happen

Label	Example	Situation entity type(s)
Habitual Event	I used to drive to work.	Generalizing Sentence
	Recycling is a good idea.	Generic Sentence
Episodic Event	Anne wrote a paper.	Event
On-going Event	Bush was reading a story.	Event
Completed Event	1 million Vietnamese refugees have been resettled.	State
State	I need a notebook.	State

Table 3.2: Event type labels of Xue and Zhang (2014) and Zhang and Xue (2014), showing the situation entity types corresponding to the respective examples. Some examples are from Xue and Zhang (2014) and the corresponding annotation manual.

on a regular basis, but this label also is used for statements that express general tense-less truths. Episodic Events describe situations that involve a change or occurrence "in a relatively short period of time"; On-going Events are usually indicated by the Progressive in English. When Completed Events occur in the Perfect in English, they are considered as States in our annotation scheme.

In addition to event type, Xue and Zhang also annotate *modality*, marking states and events as actual, intended (i.e., expected or planned), hypothetical (e.g., in conditional clauses) or modalized. In our situation entity annotation scheme, we map many of these cases to State, as they express possible states of the world.

The corpus (at the time of publication) consists of 24527 annotated text spans in 6289 sentences. After three rounds of training, three annotators, all native speakers of English, achieved an observed agreement of around 80% for event type and around 90% for modality. Xue and Zhang (2014) train a CRF model using gold eventuality type and modality as features in addition to other syntactic-semantic features, and show that they are both relevant to the prediction of tense.

Zhang and Xue (2014) take this work one step further; building on Smith and Erbaugh's (2005) observation that by default, states hold in the present but (episodic) events occur in the past. Eventuality type and modality are tied to tense, and this could be used to infer semantic tense in tense-less languages such as Chinese. Zhang and Xue investigate two ways of using this information. First, they train statistical models to predict eventuality type and modality (on the Chinese side) and then use this as features to predict tense; second, they train joint models between semantic tense and eventuality type or modality respectively. For modality labeling, they use the character string of an event, the POS tags, whether it's in a conditional clause, purpose or reason clause, and whether an event occurs right at the beginning of a sentence. For eventuality type labeling, they use the character string of the event, the POS tags, adverbs on the left that modify the event, aspect markers that follow the event and whether the event is in a relative clause. The accuracy of their maximum entropy model for labeling modality is around 76% compared to a majority class baseline of 67%, and accuracy for labeling eventuality type is around 65% with a majority class baseline of 35%.

Loáiciga and Grisot (2016) automatically predict labels for *boundedness* of verb phrases with the aim of choosing the correct French tense when translating English sentences in Simple Past. Depending on the semantics of the sentence to be translated, the correct French tense might be the Passé composé, the Imparfait, the Passé simple or the Présent. They argue that one factor for making this choice is boundedness, which relates to whether the endpoints of a situation are realized in a particular context. Boundedness is distinct from telicity (Section 2.1.2), which applies at the event type level. Telic events can be realized as *bounded* or *unbounded* in a particular context, as illustrated by (4).

(4) (a) Max ran a mile. (**bounded**)
(b) Max was running a mile. (**unbounded**)

Loáiciga and Grisot annotate a small corpus of 435 sentences as bounded or unbounded, reaching an agreement of $\kappa = 0.84$ (Cohen, 1968). Unbounded eventualities take *for*-adverbials and pass the entailment test with the Progressive (see the discussion of (19) in Section 2.1.4); Loáiciga and Grisot give "sit behind a huge desk" as an example of this class. Bounded eventualities (e.g., "write an email") take *in*-adverbials and do not pass the Progressive test. They then train a classifier on this corpus, which they use to automatically label the English side of a large parallel corpus. Using a set of syntactic and lexical features, their classifier reaches an accuracy of 82.2% (the distribution of classes in their corpus is roughly balanced). They show that using boundedness as a feature during machine translation leads to an increase in BLEU score of up to 1.6.

Automatic verb classification following Levin. Another related area involving *verb classification* addresses Levin's (1993) set of verb classes, which is based on the assumption that the syntactic behavior of verbs reflects their meaning. Levin's classification is based on the set of syntactic alternations that a verb may undergo. Example (5) by Levin (1993) illustrates the *instrument subject alternation*, which is possible for verbs like "break" (5a), but not for verbs such as "eat" (5b).

(5) (a) David broke the window with a hammer.
The hammer broke the window.
(b) Doug ate the ice cream with a spoon.
*The spoon ate the ice cream.

Levin groups verbs into about 200 classes based on their semantics, following characteristics such as whether the verb causes a change of state (e.g., "break"), whether it indicates contact ("touch") or a transfer of possession (e.g., "buy"). There is a respectable amount of work on automatically identifying Levin-style verb classes in computational linguistics; we here review only some of this literature as this kind of classification does not primarily address aspectual classification.

Levin provides an index for approximately 3000 verbs out of which 784 have more than one class. Lapata and Brew (1999) frames the automatic verb classification task as a probabilistic model based on the syntactic frames that the verb takes, e.g., NP-V-NP-NP. Frequency counts for co-occurrences of verbs and syntactic frames are obtained from BNC

(Aston and Burnard, 1998). They concentrate on a small subset of verbs that are polysemous and that take one or more of three alternations, reaching good accuracies of 80-90% for labeling them with their Levin class. Merlo and Stevenson (2001) build a supervised classifier to assign verbs that are optionally intransitive to one of three classes, depending on the syntactic alternations when comparing their transitive and intransitive uses. They only use a small set of syntactic-semantic features: transitivity, causativity, animacy, voice and part-of-speech; the first three features are based on corpus frequencies of the verb's occurrences. Joanis et al. (2007) investigate the use of a broader set of features including information on the frequency with which arguments of verbs occur in different syntactic argument positions. They also employ a wider range of tense, voice and aspect features as well as animacy information of argument noun phrases ("person" mentions as identified by the chunker they are using (Abney, 1991)). Li and Brew (2008) perform similar experiments using 48 of Levin's verb classes for which there are at least 10 verbs (Graff et al., 2003) that each occur at least 100 times in the English Gigaword Corpus. They find that a mixture of syntactic and lexical information works best. Schulte Im Walde (2006) conducts clustering experiments for a set of 883 German verbs, using a manual classification of 43 semantic verb classes similar to that of Levin.

Work in Slavic linguistics. The distinction between the perfective and imperfective aspect has been extensively studied in Slavic linguistics (Comrie, 1976). Aspect is encoded lexically in Slavic languages: most verbs are either inherently perfective or imperfective, which poses a challenge when attempting to automatically translate English sentences into, for instance, Russian or Polish (Buschbeck et al., 1991; Gawronska, 1992; Kupść, 2003). In order to choose the correct tense when translating from Russian to English, German or Turkish, Zangenfeind and Sonnenhauser (2014) propose to add aspectually relevant information to the lexemes in their rule-based machine translation system (no results have been reported to date).

In Slavic languages, perfective verbs can be derived from imperfective verbs and vice versa via prefixation, suffixation or other, partially irregular, morphological transformations. There is, however, no consensus in Slavic linguistics regarding approaches of treating aspect, e.g., scholars do not even agree on whether certain affixes are part of the lexeme or merely derivational affixes. An exhaustive review of the related work is beyond the scope of this thesis; below we mention some approaches which are of interest for future computational work.

On the linguistic side, Młynarczyk (2004) and Aalstein and Blackburn (2007) develop a system for classifying Polish verbs according to the prefixes and suffixes they take and show that the induced classes correspond to states, processes, culminating processes, unitisable processes and culminations. Samardžíc and Milićevíc (2016) propose a framework for a data-driven approach to acquiring Croatian and Serbian verb aspect. They use verb aspect matrices to represent the regular aspectual sequences that exist for a particular base verb. By analyzing all verbs occurring in a Serbian translation of the novel "1984" by George Orwell, they create a database for 834 verb types. Their data set offers a starting point for a data-driven analysis of verb aspect, quantitative linguistic studies or computational modeling.

3.2 Genericity

In this section, we first briefly describe work on automatically predicting habituality on the clause-level. We then review previously developed annotation schemes for NP-level genericity and give an overview of related work on automatically identifying generic NPs.

3.2.1 Automatic identification of habituals

Habituality, as described in Section 2.2.2, is one of the phenomena subsumed by genericity. Despite the extensive treatment of habituals in the linguistic literature, there is very little work in computional linguistics addressing their automatic recognition.

Mathew and Katz (2009) address the problem of supervised categorization for habitual versus episodic sentences. The authors randomly select 1052 sentences for 57 verbs from the Penn TreeBank (Marcus et al., 1993) and manually mark them with regard to whether they are **habitual** or **episodic**. They state that they focus on verbs that are lexically dynamic, yet, their examples (Mathew, 2009) include episodic statives such as (6) and (7).

(6) (a) She was depressed for a minute. (*episodic*)
(b) She was depressed some nights. (*habitual*)

(7) Angus Young wore a school uniform twice this week. (*habitual*)

Mathew and Katz (2009) discuss a variety of syntactic features, which they extract from gold standard parse trees. The features include tense, whether the clause has Progressive or Perfect aspect, whether there are any quantificational or specific temporal adverbs, and information on whether the subject and object are definite or bare plurals. In addition, presence of conditional and prepositional phrases is recorded. Their aim is to study the potential of using syntactic features alone to identify habitual sentences. They compare a decision tree and a Naive Bayes classifier using 10-fold cross validation on their data set. Both algorithms reach comparable performance with slightly different precision-recall trade-offs. Always assigning the majority class (episodic) would result in an average precision of 73.1%; precision and recall are around 83% and 62% for the habitual class and 87% and 95% for the episodic class respectively.

Other recent related work (Williams, 2012; Williams and Katz, 2012) extracts typical durations (in terms of actual time measures) for verb lemmas from Twitter. They distinguish episodic and habitual uses of the verbs, using the method of Mathew and Katz (2009), and collect typical durations for episodic and habitual uses separately for each verb. For example, they automatically determine that the duration of "kiss" in episodic use should be measured in "seconds," while its habitual use should be measured in "weeks."

3.2.2 Annotation of NP-level genericity

This section surveys previous approaches to annotating genericity at the NP level. We first give an overview of the ACE corpora, which have been the most widely used for recent research on automatically identifying generic NPs (Reiter and Frank, 2010), and then explain other approaches.

ACE entity class annotations. The research objective of the Automatic Content Extraction (ACE) program (1999-2008) was the detection and characterization of entities, relations and events in natural text (Doddington et al., 2004). ACE-2 (Mitchell et al., 2003) and ACE-2005 (Walker et al., 2006) are the two most notable annotation projects for labeling genericity of NPs to date. All entity mentions receive an *entity class* label indicating their genericity status.

In the ACE-2 corpus, 40106 entity mentions in 520 newswire and broadcast documents are marked with regard to whether they refer to "any member of the set in question" (GEN, generic) rather than "some particular, identifiable member of that set" (SPC, specific/non-generic).[4] This leads to a mix of constructions being marked as generic as illustrated by example (8): types of entities (a), generalizations across a set of entities (b), hypothetical entities (c) and negated mentions (d).

(8) (a) Good students do all the reading. (GEN)
(b) Purple houses are really ugly. (GEN)
(c) If a person steps over the line, ... (GEN)
(d) I saw no one. (GEN)

Suggested attributes of entities are marked as generic (9a), but a "positive assertion test" leads to marking both NPs ("Joe" and "a nice guy") as specific in examples like (9b). Neither of these two cases ("be a nice person" / "be a nice guy") is in fact an entity mention; they are rather predicative uses.

(9) (a) John seems to be a nice person. (SPC, GEN)
(b) Joe is a nice guy. (SPC, SPC)

In addition, in both ACE-2 and ACE-2005, modifier uses of nouns, to which the genericity distinction is not applicable, also receive labels (e.g., "a subway system"). The major drawback of ACE-2 is that genericity is basically defined as lack of specificity, which leads to uncertainty and inconsistencies in the annotation process, and to a heterogeneous set of NPs labeled with GEN, including quantificational NPs and NPs in modalized, future, conditional, hypothetical, negated, uncertain, and question contexts.

The guidelines for genericity were redefined for annotation of the **ACE-2005 Multilingual Training Corpus** (Walker et al., 2006), which contains news, broadcast news, broadcast conversation, forum and weblog texts as well as transcribed conversational telephone speech. In contrast to ACE-2, the ACE-2005 annotation manual[5] clearly defines mentions as kind-referring or not, using the labels GEN (generic) and SPC (specific/non-generic) respectively. In the updated guidelines of ACE-2005, the label USP (underspecified) is introduced for non-generic non-specific reference as in (10a/b). Moreover, annotators are

[4] See "Entity Detection Tracking and Metonymy Annotation Guidelines, Version 2.5.1", available from LDC: https://catalog.ldc.upenn.edu/docs/LDC2003T11/

[5] See "ACE English Annotation Guidelines for Entities, Version 5.6.6" (available from LDC) or 2008's version 6.6.

asked to mark truly ambiguous cases that have both a generic and a non-generic reading as USP (10c).

(10) (a) Many people will participate in the parade. (USP)
(b) We will elect five new officials. (USP)
(c) The economic boom is providing new opportunities for women in New Delhi. (USP)

The class also contains mentions of an entity whose identity would be "difficult to locate" ("Officials reported ..."). In our opinion, the latter interferes with the definition of SPC as marking cases where the entity referred to is a particular object in the real world, even if the author does not know its identity (11). The breadth of the USP category causes problems with consistency of application (see Section 8.4.2).

(11) At least four people were injured. (SPC)

While we agree that in general there are underspecified cases, the guidelines for ACE-2005 mix other phenomena into the USP class, resulting in a high confusion between USP and SPC, as well as USP and GEN, in the manual annotations (Friedrich et al., 2015b). Data from two annotators is available, and we compute an agreement of Cohen's $\kappa = 0.53$ over the four labels. The ACE corpora consist only of news data, and the distributions of labels are highly skewed towards specific mentions. For some criticism of the ACE annotation scheme, see also the work of Suh (2006).

The ACE annotation scheme has also been applied in the **Newsreader** project.[6] The ECB+ corpus (Cybulska and Vossen, 2014) is an extension of EventCorefBank (ECB), a corpus of news articles marked with event coreference information (Bejan and Harabagiu, 2010). ECB+ annotates entity mentions according to ACE-2005, but collapses the three non-GEN labels into a single category. Roughly 12500 event participant mentions are annotated, some doubly and some singly. Agreement statistics for genericity are not reported.

Agreement in the ACE corpora. The ACE corpora were first labeled by two annotators independently, then adjudicated by a senior annotator. To our knowledge, agreement numbers on this task have not been published to date. In order to assess both the quality of the data and the difficulty of the task, we compute inter-annotator agreement as follows. Using the 533 documents from the adjudicated data set that were marked by two annotators in the first step, we compute Cohen's κ (Cohen, 1968) for entity class annotations over the four labels SPC, GEN, USP and NEG.

Intuitions about NP genericity are most reliable for subject position as other argument positions involve additional difficulties (Link, 1995).[7] To get a better sense of the difficulty

[6] `www.newsreader-project.eu`

[7] Link (1995) discusses the phenomenon of *dependent generics*: an example of a dependent generic NP is "manes" in the sentence "Lions have manes." The semantics of this sentence is "for every typical lion **x**, there is a mane which **x** has." The NP "manes" does not refer to a kind per se: there is no "kind mane" that the "kind lion" has. Such dependent generics are not easily recognizable based on their syntax, e.g., in the sentence "The leopard has a close relative, the black panther," the object NP also refers to a kind.

of annotating subjects compared to that for other argument positions, we compute agreement over mentions whose (manually marked) head is the grammatical subject of some other node in a dependency graph. We obtain dependency graphs using the Stanford parser (Klein and Manning, 2002) and identify subjects by considering any dependency relation matching the pattern `*subj`.

An additional complication in entity mention annotation is determining the mention span. Because spans are not pre-marked in the ACE corpora but identified independently by each annotator, we compute κ only over all exactly-matching entity mention spans for the two annotators. For all mentions, annotators mark about 90% of spans marked by the other annotator. For subject mentions, this number is even higher, at about 95%. The spans of the remaining mentions overlap for the two annotators. We exclude them from this study as we cannot be sure that the two mention spans refer to the same entity.

Table 3.3 shows the confusion matrices of labels for the all-mentions-case and the subjects-only case. In both cases, confusion between SPC and GEN is acceptable, but confusion between USP and both SPC and GEN is rather high. For example, in the case of subjects, annotator 1 tags 652 mentions as GEN that annotator 2 marks USP, but the two of them only agree on 597 mentions to be GEN. Although it may be useful to create a separate category for unclear or underspecified cases, the definition of USP is not yet clear-cut and compounded with lack of *specificity*, which refers to whether the speaker presumably knows the referent's identity or not. Even if the identity of a referent may be 'difficult to locate' (as in "Officials reported ..."), the clause certainly does not make a statement about the *kind* 'official'; instead, it expresses an existential statement ("There are officials who reported ..."). The definition of SPC states that the reader does not necessarily have to know the identity of the entity, possibly making the distinction hard for annotators.

Another difficult case are noun modifiers in compounds (e.g. "a subway system"); these are marked as GEN in the corpus. Using the automatic parses, we find that 9.5% of all mentions marked GEN in the adjudicated corpus are one-token mentions modifying another noun via an *nn* dependency relation. Genericity as reference to kinds is an attribute of referring expressions, which, in most cases, cannot be determined without interpreting the surrounding discourse. Because nominal modifiers do *not* introduce discourse referents, they should not be treated on the genericity annotation layer.

The data shows moderate agreement for the first two passes of entity class annotation ($\kappa = 0.53$ for all mentions and $\kappa = 0.50$ for subject mentions). Note that κ scores are not directly comparable across different annotation projects. We give the above scores for the sake of completeness. Observed and expected agreement are 0.83 and 0.65 for the all-mentions case and 0.79 and 0.58 for subject mentions. This indicates that the all-mentions case may contain some trivial cases, one of which is the case of nominal modifiers described above.

In summary, the ACE scheme problematically fails to treat subject NPs differently from NPs in other syntactic positions, and 'fuzzy' points in the guidelines, particularly concerning the USP label, contribute to disagreements between annotators.

All mentions		Annotator 2			
		SPC	USP	GEN	NEG
Annotator 1	SPC	28168	1575	684	3
	USP	1142	1954	963	2
	GEN	757	1261	1707	10
	NEG	8	5	7	71

Subjects only		Annotator 2			
		SPC	USP	GEN	NEG
Annotator 1	SPC	9830	830	234	1
	USP	634	1091	476	1
	GEN	272	652	597	4
	NEG	4	1	2	46

Table 3.3: Confusion matrices of entity class tags for ACE 2005 for mentions where annotators agree on spans.

Other corpora annotated at the NP-level. The resources surveyed here apply carefully-defined notions of genericity but are too small to be feasible machine learning training data.

The question of whether an NP is generic or not arises in the research context of coreference resolution. Some approaches mark coreference only for non-generic mentions (Hovy et al., 2006; Hinrichs et al., 2004); others include generic mentions (Poesio, 2004), or take care not to mix coreference chains between generic and non-generic mentions (Björkenstam, 2013). Björkelund et al. (2014) mark genericity in a corpus of German with both coreference and information-status annotations. Nedoluzhko (2013) surveys the treatment of genericity phenomena within coreference resolution research; they provide a complete overview. In short, they argue that a consistent definition of genericity is lacking and report on their annotation scheme for Czech as applied to the Prague Dependency TreeBank (Böhmová et al., 2003).

The **GNOME corpus** (Poesio, 2004) is a coreference corpus with genericity annotations; NPs are marked with the attributes `generic-yes` or `generic-no`. Poesio reports that the annotators found it hard to decide how to mark references to substances ("A table made of wood") and quantified NPs. Similar to our experience, he found it helpful to have annotators first try to identify generic sentences, and then determine this attribute of the NP. He reports an agreement of $\kappa = 0.82$ on the GNOME corpus, which consists of 900 finite clauses from descriptions of museum objects, pharmaceutical leaflets and dialogues.

Coming from a formal semantic perspective, Herbelot and Copestake (2009, 2010, 2011) describe an approach to treating **ambiguously quantified NPs**. This annotation effort aims to produce resources for the task of determining the extent to which the semantic properties ascribed to a given NP in context apply to the members of that class. For example, the statement "Cats are mammals" describes a property of *all* cats, while "Cats have four legs" is true only for most cats. The scheme, which includes the labels One, Some, Most, All and Quant (for explicitly quantified NPs), is applied to 300 subject-verb-object triples from sentences randomly extracted from Wikipedia. Annotators are shown the sentence and the triple. κ ranges from 0.88 and 0.81 for Quant and One to values between 0.44 and 0.51 for the other classes. Using this corpus, Herbelot and Copestake (2011) conduct first experiments on identifying the quantifier for an NP. They train a decision tree using article and number information for the NP, as well as the tense of the verb following it. Overall precision of their classifier is 78%; F-scores range from 10% (Most) to 92% (One).

Bhatia et al. (2014b) present an annotation scheme for **Communicative Functions of Definiteness**, intended to cover the many semantic and pragmatic functions conveyed by choices regarding definiteness across languages of the world. They annotate English, Hindi, Hebrew and Russian texts. The scheme has been applied to 3422 English NPs contained in texts from four genres. Their typology includes two categories relevant to genericity: Generic_Kind_Level applies to utterances predicating over an entire class, like "Dinosaurs are extinct." Generic_Individual_Level is for predications applying to the individual members of a class or kind, such as "Cats have fur." Across 1202 NPs annotated for an inter-annotator agreement study, the two annotators used the Generic_Individual_Level label 45 times and 30 times, respectively, with agreement in 29 cases. Neither used the Generic_Kind_Level. The entire corpus contains just 131 NPs labeled with Generic_Individual_Level and none with Generic_Kind_Level. Bhatia et al. (2014a) train a log-linear model and a decision tree for predicting the communicative function of the NPs in their corpus, using a variety of syntactic-semantic features capturing properties of the target NP as well as the verbs and NPs in its immediate context. The majority baseline when evaluating according to exact label match is 12.1; accuracy of the decision tree is 49.7%. Bhatia et al. also evaluate according to a soft match measure, which gives partial credit when the predicted label is related to the correct one. In this setting, the majority baseline accuracy is 47.8% and the log-linear model achieves the best results of 78.2%. Precision and recall for the Generic_Individual_Level class are 14 and 11% respectively.

The question of genericity has also been addressed in cognitive science (Prasada, 2000). Gelman and Tardif (1998) study the usage of generic NPs cross-linguistically for English and Chinese in child-directed speech. They annotate kind-referring NPs as generic. They report agreement as the fraction of items on which the annotators agreed at over 99%, but given that their data set has fewer than 1% generic NPs, this statistic does not allow us to estimate how well annotators agreed.

3.2.3 Automatic identification of reference to kinds

With the aim of extracting common sense knowledge from the web for ontology building, Suh et al. (2006) propose a rule-based approach for extracting generic NPs. Their approach extracts only bare plurals and singular NPs quantified with "every" or "any" as generic. They evaluate precision against the entity class annotations in the ACE-2005 corpus, reaching 34% for SPC, 29% for GEN and 37% for USP.

Reiter and Frank (2010) use a wide range of syntactic and semantic features to train a supervised classifier for identifying generic NPs. Their features capture properties of the NP itself as well as the sentence that includes it. Syntactic features include part-of-speech tags, dependencies, tense, voice, mood, etc.; semantic features include lemma and WordNet information. Reiter and Frank's set of features is a major inspiration for part of the features implemented in our system as described in Section 7.1. They train and evaluate a Bayesian network using 10-fold cross validation on the ACE-2 corpus, labeling NPs as generic or non-generic. The majority class baseline reaches an accuracy of 86.8%, a baseline only relying on the person information for each NP reaches an accuracy of 87.2% and a macro-average F1-score of 62.9%. Reiter and Frank's system has an accuracy of 80.6% and a macro-average F1-score of 71.3%. They find number, person, part-of-speech, determiner type, bare plural, dependency relations, the NP's WordNet sense, tense and the main verb's lemma to be the most informative among their features. This work is a competitive baseline for our experiments presented in Section 8.4.

The work of Palmer et al. (2007), described in detail in Section 3.1.4, is also related to automatic approaches to identifying generic sentences. They classify clauses into several types of situation entities including states, events, generalizing sentences (habitual utterances referring to specific individuals) and generic sentences (see also Section 3.1.4). However, generic sentences are extremely sparse in their data set.

3.3 Other related work

In this section, we give an overview of further work that is related to ours from various perspectives.

General vs. specific sentences. Louis and Nenkova (2011) describe a method for automatic classification of sentences as *general* or *specific*. *General* sentences are loosely defined as those which make "broad statements about a topic," while *specific* sentences convey more detailed information. This distinction is not immediately related to the phenomena treated as generics in the literature. Kind-referring subjects can occur in both *general* (12) and *specific* (13) sentences; *general* sentences can also have non-kind-referring subjects (14).

(12) Climatologists and policy makers [...] need to ponder such complexities [...]. (**general**)

(13) Solid silicon compounds are already familiar – as rocks, glass, [...]. (**specific**)

(14) A handful of serious attempts have been made to eliminate ... diseases. (**general**)

The authors use a proxy for annotated data by extracting pairs of sentences from the Penn Discourse Treebank (Prasad et al., 2008) which are related by one of two relevant discourse relations: Instantiation or Specification. The two groups of sentences are then treated as training instances for the generic and specific classes, disregarding their original pairwise association as arguments to a single instance of a discourse relation. The classifier trained on PDTB sentences is then evaluated on a smaller corpus annotated via crowd-sourcing and described in more detail by Louis and Nenkova (2012). The corpus sentences are from three types of news articles, each is marked by five different annotators, and agreement is reported as high, especially considering the rather intuitive instructions given to annotators. A logistic regression classifier is trained using features such as sentence length, polarity of opinion words, WordNet specificity, words, language model likelihood and counts of particular syntactic constructions. Accuracy of the classifier is 75.9% for the part of the data set extracted from Instantiation relations and 59.5% for the part extracted from Specification relations. Accuracy for 10-fold cross validation on the crowd-sourced corpus is close to 80%.

Discourse type. Cocco et al. (2011) aim to produce a categorization of sequences of text according to Adam's (2011) *séquences élémentaires prototypique (text sequence prototypes)*. Adam distinguishes the five types *narrativ, descriptif, argumentatif, explicatif* and *dialogal*. Cocco et al. (2011) cluster clauses of a text into discourse types and investigate the distributions of part-of-speech (POS) tags depending on the type of text. A text corpus of three 19th century French short stories by Maupassant has been clause-segmented and annotated by a human expert. The corpus contains 504 sentences and 905 clauses. The clauses are labeled with five types as suggested by Adam and a sixth type *injunctive* following Bronckart (1997). The injunctive type covers instructions and incentives for actions. They assign POS tags to the words in each clause, and then cluster the clauses based on their distributions of POS tags, comparing the standard version of K-means to a fuzzy variant of K-means. They find adjectives to be relevant for the descriptive type, simple past tense for the narrative type and future tense for the type dialogal. While they find a relation between the automatic clustering and the expert classification of types, the automatic clustering can by no means be used to label clauses: accuracy would be very low.

Cocco (2012) is a continuation of the above mentioned research using the same data (plus one more document, resulting in a total of 764 sentences and 1087 clauses) but more complex feature spaces. In this work, they obtain distributions over POS-tag n-grams (for uni-, bi- and trigrams) for each clause. For the clustering algorithms, they use chi-squared distances between clauses, applying some power transformations to the n-gram counts. The number of groups for K-means is here set to 6 (the number of discourse types from above). For this clustering task, they do not find bi- and tri-grams to improve over simply using POS-tag unigrams and attribute this to the sparsity problem. However, using high-dimensional embeddings of the features seems to improve results.

Further clause-level classifications. Finally, our work is related, though not closely, to research on classifying clause types based on their propositional content or function in a discourse. Teufel et al. (1999) describe a first attempt to annotate **argumentative zones** in scientific papers from the field of computational linguistics. In their corpus, 12,783 sentences are marked according to their rhetorical status. The distribution of argumentative zones is skewed (67% Own, 16% Other, 6% Background, 5% Contrast, 2% Basis, 2% Textual and 2% Aim). Based on this corpus, Teufel and Moens (2002) propose a supervised approach to labeling sentences with their rhetorical status (see also Teufel, 1999; Teufel and Moens, 2000). Their system relies on features representing the sentence's location in the document and paragraph, sentence length, whether the sentence contains 'important terms' (as determined by tf.idf) and whether it contains citations or self-citations. Verb syntax of the first finite verb in the sentence is captured by features presenting voice, tense and whether it is modalized. They manually design *meta-discourse features*, i.e., patterns of formulaic expressions (such as *to our knowledge* or *in this paper*) and patterns capturing types of agents (subjects) or types of actions (the related predicates). Types of actions are for instance ARGUMENTATION (*argue, disagree*) or PRESENTATION (*present, report*). Discourse context is leveraged by using the most probable *previous* category as a feature. Using 10-fold cross validation, their naïve Bayes classifier achieves an accuracy of 73%, and a macro-average F-measure of 0.5, with a random baseline based on the categories' distribution reaching 67% and .11 respectively, and an upper bound (of comparing humans' annotations against each other) of 87% and 0.69. Guo et al. (2011) propose a weakly-supervised approach to the same problem. Using *active SVM with self-training*, using only 10% of the labeled data, they are able to outperform a supervised classifier trained on their entire data set (Guo et al., 2010), reaching 81% accuracy. Their method iteratively adds examples in an active learning setting, and uses some additional training/testing steps.

Séaghdha and Teufel (2014) propose an unsupervised approach to labeling a text with argumentative zones. Their Bayesian model incorporates the following intuitions: (a) lexically similar sentences have similar purposes, (b) sentences with the same rhetorical function are often grouped together into blocks, and (c) the words and linguistic constructs used to convey rhetorical function are independent of the paper's topical content. Their model assumes that each word is either generated from a (content) topic model or from an "un-topic" model capturing all the words belonging to the conventional rhetorical language (such as *result*, *suggest*, *method* or *significantly*). For training and evaluation of their approach, they use a set of abstracts from the domain of semi-automated cancer risk assessment, which is annotated with argumentative zones (Guo et al., 2011), as well as a large set of unlabeled and topically heterogeneous abstracts collected from open-access journals. When incorporating the zone index assigned by their unsupervised model into a supervised model as an additional feature, accuracy improves by one (significant) percentage point.

Stede and Peldszus (2012) suggest that discourse analysis would benefit from an investigation of the *illocutionary status* of discourse segments. Performing an *illocutionary act* is orginally understood as performing a *speech act* (Austin, 1975): by uttering the question "Is there any wine left?", one also performs the illocutionary act of requesting wine. Stede and Peldszus use the term in a wide sense, also covering the pragmatic role of discourse segments that would not be considered as speech acts in their own right. They define an inventory of illocution types, which comprises **Report** (situations or actions where the speaker is not actively involved), **Report-Author** (situations or actions where the author is involved), **Idents** (segments conveying feelings or desires), **Evaluatives** (judgments), **Estimates** (assumptions or prognosis), **Commitments** (where the author commits to something), **Directives** (where the author requests the reader to perform some action) and **Present-Hypothetic-Situation** (the description of some unrealized situation). Given a written annotation manual, otherwise untrained annotators mark 2350 segments that are related by some causal connective in a corpus of 250 German hotel reviews, achieving an agreement of $\kappa = 0.51$.

Part II

Corpus: annotation and agreement

Chapter 4

Segmentation: what invokes a situation?

This chapter discusses the segmentation of texts into smaller units, which we call *situation segments*, for the purpose of manual and automatic situation entity type annotation. The main questions addressed here are (a) which syntactic structures map to a situation segment, i.e., introduce situation entities to the discourse, and (b) how we approximate this segmentation in an automatic way.

Smith (2003, p. 23) suggests that situation entities are introduced by the clauses of a text. Manual segmentation of text into clauses is not a trivial task, and requires the careful design of annotation guidelines. It has been approached both from a syntactic perspective (Bies et al., 1995) and a discourse perspective (Polanyi, 1995; Carlson et al., 2001; Polanyi et al., 2004; Prasad et al., 2007). Automatic clause segmentation is also non-trivial (Tjong et al., 2001; Soricut and Marcu, 2003; Tofiloski et al., 2009). As Carlson et al. (2001) put it, "the boundary between discourse and syntax can be very blurry," and in fact clause and discourse segmentation constitute their own research areas. For this reason, we do not attempt to develop our own situation segmentation method, but approximate situation segmentation using an existing discourse segmenter based on Rhetorical Structure Theory (RST) (Soricut and Marcu, 2003).

We perform annotation of situation entities on texts that are automatically pre-segmented, which means that the annotation task is reduced to assigning labels for situation entity types and the related features to each segment. Automatic segmentation has several advantages over allowing annotators to define segments while annotating. First, the segmentation step is fully reproducible. Second, automatic segmentation enables the completely automatic processing of unlabeled texts with the same segmentation method underlying the manual annotation of the data used for training. Having all annotators label the same set of segments also greatly simplifies the computation of inter-annotator agreement on the situation entity types and features as well as the creation of a gold standard via majority voting. Both tasks would be difficult if deviations in segmentation were to be taken into account. This approach is similar to a previous approach to annotating situation entity types (Palmer et al., 2007), in which one annotator manually pre-segmented texts with the aim of avoiding segmentation-based disagreements.

In section 4.1 we give a definition of which syntactic structures we consider to introduce situation entities to the discourse; section 4.2 links our definition of situation segments to existing notions of *discourse segments*; section 4.3 explains the technical details of our automatic segmentation method and section 4.4 describes how annotators are instructed to handle the pre-segmented texts in case an automatically created segment does not invoke a situation according to their judgment.

4.1 Situation segmentation

As a preparation for manual annotation of a text with situation entity types, we automatically split the text into *situation segments*. In this section, we discuss which syntactic structures we consider to introduce situation entities to the discourse for this purpose.

According to Smith (2003, p. 23), situation entities are introduced by the clauses of a text, while noun phrases introduce individuals (e.g., people, places, objects or ideas) and tense and time adverbs introduce times. The clause's *verb constellation*, i.e., the main verb and its arguments, invokes a situation entity – a State, an Event, a General Stative or an Abstract Entity. Smith does not further specify which linguistic structures she considers to be a clause. We approximate the situation segmentation problem from a practical perspective, aiming at a set of syntactic structures for which annotators can easily provide labels, excluding some syntactic structures which are less clear.

In brief, we consider all finite clauses to invoke situation entities, as well as postmodifications of noun phrases using the present or past participle, which can be regarded as reduced relative clauses. The latter constructions are roughly equivalent to finite clauses. Example (1) illustrates the segmentation of a longer sentence according to this definition.

(1) [Among these are rules] [governing emissions] [that limit visibility in the national parks] [and rules governing pollution] [that drifts eastward from Midwestern power plants.] `(MASC news 20020731-nyt.txt)`

There are five situation segments in (1). Each finite clause is a situation segment, and the two reduced relative clauses containing the present participle "governing" are also treated as separate segments since their non-reduced version "rules that govern" clearly introduces a situation entity to the discourse. Though other verb forms such as infinitives and nominal constructions might invoke situation entities in certain cases, we do not address them here for the sake of feasibility.

As a consequence, we mark example (2) as one Event, regarding "to come" as a specification of the event introduced by "invited" rather than a separate event.

(2) Fidel Castro invited John Paul to come for a reason.
`(TimeBank ABC19980120.1830.0957.sgm)`

Perception verbs that embed an infinitive construction sometimes imply that the embedded event actually happened as in example (3). Such infinitive constructions in perception contexts are treated in the same way as other infinitives, i.e., we do not label them as situation entities.

(3) John saw Mary run.

Background. In Smith's (Smith, 2003) theory of discourse, situation entities are introduced by the "verb constellations of the main verbs of clauses." There is, however, neither a universal definition of what constitutes an event, a state or a situation in general nor a clear-cut linguistic definition of a clause. In the following, we motivate our above operationally-motivated mapping between syntactic structures and situation entities and explain links to related work.

Polanyi (1995), in an endeavor to understand the nature of the atomic units of discourse, argues that no particular linguistic structure (such as the sentence, clause, prosodic unit or paragraph) corresponds to a *minimum unit of meaning*, which contains information about exactly one event, event-type or state of affairs. Her Discourse Constituent Units (DCU), the elementary units from which tree-like discourse structures are built in the Linguistic Discourse Model (LDM) are semantic rather than syntactic. While this is theoretically appealing, a method for automatic discourse parsing relying on the current stack of natural language processing technologies will have trouble going from semantics to syntax: a method of *approximating* situation segments from syntactic structures is required.

In addition, the question of what exactly a minimum unit of meaning is remains open. Reconsider example (2): does "to come" invoke a separate future event, or is there only one event of "inviting-to-come", i.e., is "to come" merely a specification of the Event "to invite"? The answer to this question will to some extent always be up to the reader's interpretation unless a particular predefined ontology of eventuality types is assumed. For example, the annotation scheme used in the ACE 2005 Multilingual Training Corpus (Walker et al., 2006) defines a range of *Event types* and *subtypes* including BE-BORN, MARRY, DIVORCE, INJURE, DIE, MOVEMENT, TRANSPORT and TRANSFER-MONEY. Event recognition in text reduces to assigning one of the predefined Event types in this case. Creating an exhaustive set (ontology) of eventuality types in this way that would allow to cover arbitrary texts is not a realistic assumption.

In this work, we do not assume a particular predefined ontology of eventuality types. Our modeling of situation entity types aims at capturing aspectual properties of a discourse's situation entities rather than trying to map the discourse's content to a particular set of eventuality types.

TimeML (Pustejovsky et al., 2003a) is similar to our approach in this sense, but takes a less restrictive approach to annotating event structure, to some extent semantically motivated in a way similar to the approach suggested by Polanyi (1995). In TimeML, "events that happen or occur," and "states and circumstances in which something holds" are all marked as EVENTs, and they can be expressed by tensed or untensed verbs, nominalizations (4a), adjectives, predicative clauses or even prepositional phrases (4b).[1]

(4) (a) Israel will ask the United States to delay a military strike against Iraq until the Jewish state is fully prepared for a possible Iraqi attack.
(b) All 75 people on board the Aeroflot Airbus died.

Asher (1993) investigates how eventualities (states and events) and abstract entities (propositions, properties, states of affairs and facts) are referred to in natural language.

[1] Examples by Saurí et al. (2005b).

He provides an inventory of *sentential nominals*, i.e., syntactic structures whose meanings are correlated with sentences. We compare our mapping between syntactic structures and situation segments to this inventory with the aim of highlighting which non-finite and nominal event-denoting constructions fall outside of our definition of situation-invoking syntactic structures. The inventory of *sentential nominals* (Asher, 1993, p. 138) contains the following syntactic constructions: derived nominals (5), gerund phrases (6), *that*-clauses (7), *for*-infinitival phrases (8), naked infinitive phrases (3a) and noun phrases involving common nouns that may combine with *that*-clauses or gerund phrases (9).[2]

(5) (a) The army's destruction of the city
(b) Franklin's favorite invention

(6) (a) The mayor's throwing of the pizza
(b) John's hitting Bill
(c) The gathering of the pecans in Central Texas

(7) that Sam greeted Susan

(8) (a) For Fred to shoot Bill is not something I desire.
(b) John wanted for Mary to be chair.

(9) (a) Mary's doubt that John was unhappy
(b) The fact that John was unhappy
(c) The letter explaining the situation

Some of these constructions such as (5a) or (7) clearly refer to events while others may also refer to objects (5b). Of the above sentential nominals, we mark only *that*-clauses as situation entities. We do not treat the rest of the above sentential nominals as invoking situations, but simply as part of the larger situation segment in which they are embedded. We decided not to mark them because the boundary between event-denoting constructions (5a) on the one hand and phrases denoting concrete (5b) or abstract objects (9) on the other hand is not clear; additional annotation guidelines would be necessary.

4.2 Discourse segmentation and situation segmentation

Discourse segmentation, the task of splitting a text into units that are related by various discourse or coherence relations is closely related to the task of situation segmentation. We find discourse segmentation as applied in Rhetorical Structure Theory (RST) to be most similar to our segmentation and use an existing discourse segmenter as the basis for our automatic segmentation of texts into situation segments. Here we discuss the relationship between the two tasks.

As explained above, our specification of situation segmentation is operationally-motivated, capturing a subset of syntactic structures that clearly introduce situation entities to the discourse. The primary question in discourse segmentation, as a prerequisite

[2] Examples by Asher (1993).

for automatic discourse parsing, is the question of whether other discourse units can attach to the segment. As a consequence, the various approaches to discourse segmentation are to some extent specific to the respective theory of discourse structure. The definitions of what constitutes a discourse unit depend on the inventory of discourse and coherence relations of the respective theory. Similarly, our definition of situation-invoking syntactic structures is based on the inventory of situation entity types by Smith (2003).

Segmentation in the PDTB. In the Penn Discourse Treebank (PDTB), implicit relations occur between adjacent sentences by definition, and explicit relations may relate subordinating or complement clauses and nominalizations (10) as well (Prasad et al., 2007, 2008).

Example (10) illustrates a difference in PDTB and situation segmentation. The discourse relation between ARG1 and ARG2 is marked as Contingency.Condition.Hypothetical, indicating a relationship between a hypothetical scenario in ARG1 and a possible consequence in ARG2 (Prasad et al., 2007, p. 26). The PDTB arguments include the phrase "for major new liberalizations", denoting a possible event. This prepositional phrase is not considered to invoke a situation entity according to our guidelines.

(10) **PDTB segments:** ... and many are hoping [for major new liberalizations]$_{ARG2}$ if [he is returned firmly to power.]$_{ARG1}$
Situation segments: ... [and many are hoping for major new liberalizations] [if he is returned firmly to power.]

In general, segmentation in the PDTB is often coarser-grained than the segmentation of discourse into situation segments. In PDTB, in the extreme case, if an argument of a connective is an abstract object, it can even be realized as multiple sentences as shown in example (11). The same paragraph would be segmented into multiple situation segments.

(11) **PDTB segmentation:** [Here in this new center for Japanese assembly plants just across the border from San Diego, turnover is dizzying, infrastructure shoddy, bureaucracy intense. Even after-hours drag; "karaoke" bars, where Japanese revelers sing over recorded music, are prohibited by Mexico's powerful musicians union.]$_{ARG2}$ Still, [20 Japanese companies, including giants such as Sanyo Industries Corp., Matsushita Electronics Components Corp. and Sony Corp. have set up shop in the state of Northern Baja California.]$_{ARG1}$

Segmentation in the LDM. In the LDM (Polanyi, 1995), the atomic units of discourse are called Basic Discourse Units (BDUs); they are segments that have the potential to establish an anchor point for future attachment of other segments. Polanyi et al. (2004) start by defining the semantic basis for functioning as a segment and then identify syntactic constructions that are able to carry the semantic information needed for discourse segment status. They observe that in written text, often a subsequent but not necessarily adjacent segment continues the development of material introduced in a sub-sentential, often subordinate, clause. In the following, we discuss some examples from Polanyi et al. (2004) with regard to whether their BDUs correspond to situation segments.

(12) **LDM segments:** [California elected Schwarzenegger] [governor].
Situation segments: [California elected Schwarzenegger governor.]

For the purpose of situation segmentation, we treat the entire sentence in (12) as one situation entity: small clauses are not considered to invoke separate situation entities because they do not contain a verb. In example (13), BDU and situation segmentation are the same, treating the post-modifier *braying next door* as a separate situation, as it can be interpreted as *who was braying next door*. We do not require situation segments to be continuous in the text.

(13) **LDM/Situation segments:** [The donkey [braying next door] was annoying.]

Polanyi et al. also give examples for non-BDU segments. The following nominal gerunds and nominalizations are non-BDU segments. We do not consider them to be situation segments either.

(14) **LDM/Situation segments:** Singing is fun. *(gerund)*

(15) **LDM/Situation segments:** Rationalization is useless. *(nominalization)*

Some nominals may refer to specific events as in example (16). However, many do not – as in (14) – and as explained above, we leave methods for distinguishing these two types and annotating gerunds for future work.

(16) **LDM/Situation segments:** The destruction of the old town hall really was a big loss for our city.

Segmentation in RST. In RST, the size of Elementary Discourse Units (EDU) is in principle arbitrary, but the units should have independent functional integrity (Mann and Thompson, 1988). In the original analyses, Mann and Thompson state that units are essentially clauses, but clausal subjects, complements and restrictive relative clauses are considered as parts of the clause headed by their governing verb.

When building the RST Discourse Treebank, Carlson et al. (2001) note that applying this intuitive notion is difficult when aiming for a large and consistently annotated corpus. They develop an extensive set of rules for identifying EDUs based on syntactic constituents with the aim of obtaining a balance between tagging granularity and the ability to identify units consistently (Carlson and Marcu, 2001, section 2 and Appendix I).

We here compare their rules to our definition of which syntactic structures invoke situation entities.[3] In RST, main clauses (17) and subordinate clauses with discourse cues

[3] Examples by Carlson and Marcu (2001); Carlson et al. (2001).

(18) are considered to be EDUs, both cases corresponding to our definition of situation segments.

(17) **RST/Situation segments:** [The company will shut down its plant.]

(18) **RST/Situation segments:** ... [although it will not dismiss any employees.]

Clausal subjects, clausal objects and clausal complements are not treated as EDUs, with the exception of complements of attribution verbs (19). If a clausal subject is in fact a subordinate clause, such as the relative clause in example (20), it constitutes its own EDU.[4]

(19) **RST/Situation segments:** [The company **says**] [it will shut down its plant.]

(20) **RST segments:** [Under Superfund, those] [who owned, generated or transported hazardous waste] [are liable for its cleanup...]
Situation segments: [Under Superfund, those who owned, generated or transported hazardous waste] [are liable for its cleanup...] (wsj1331)

(21) **RST/Situation segments:** [**Making computers smaller** often means **sacrificing memory**.] (wsj2387)

For these cases, situation segmentation is mostly parallel to segmentation in RST. We do not mark the bold phrases in example (21) as situation segments because nominal gerunds are an unclear case with regard to whether they invoke situation entities or not. Situation segmentation involving subordinate clauses is also the same as in RST: finite clauses constitute their own segments (22) while infinite complements do not (23).

(22) **RST/Situation segments:** [The company announced] [that it will shut down its plant] [and dismiss several hundred employees.]

(23) **RST/Situation segments:** [The company plans to shut down its plant and dismiss several hunded employees.]

In both RST and situation segmentation, relative clauses, nominal postmodifiers as well as temporal clauses (24) are treated as separate segments.

(24) **RST/Situation segments:** [Just months before dismissing several hundred employees,] ...

In the RST Discourse Treebank (Carlson and Marcu, 2001; Carlson et al., 2001), phrasal expressions are allowed as discourse units if they begin with one of a finite set of discourse cues, e.g., "**as a result of** margin calls." As stated above, we do not mark nominal constructions with situation entity types.

The set of rules by (Carlson and Marcu, 2001) are motivated by RST's inventory of discourse relations (schemas). While infinitival constructions are generally not considered

[4] Example by Carlson and Marcu (2001).

to constitute their own EDU, they are if they introduce a purpose clause as in (25) because the infinitival clause corresponds to the satellite of a Purpose relation here. Prepositional phrases with clausal objects (26) are EDUs, while other non-finite clausal objects are not. We do not make these distinctions in situation segmentation, and treat each of the two above cases as a single segment. In general, the EDU segmentation of RST closely corresponds to our definition of situation segmentation while being more fine-grained in certain cases.

(25) **RST:** [... officials at Southern Co. conspired to cover up their accounting for spare parts] [**to evade** federal income taxes.]
Situations: [... officials at Southern Co. conspired to cover up their accounting for spare parts to evade federal income taxes.]

(26) **RST:** [Canadian Utilities isn't alone] [in exploring power generation opportunities in Britain.]
Situations: [Canadian Utilities isn't alone in exploring power generation opportunities in Britain.]

Summary. To sum up, there is no general agreement of what constitutes a discourse segment. The various definitions depend on the granularity and inventory of the underlying theory of discourse structure, and they correspond to our definition of a situation segment to varying degrees. We found the greatest overlap between the notions of discourse segments and situation segments in the case of RST. In the next section we move away from these theoretical concerns to a practical solution for automatically generating a segmentation of texts into situation segments, making use of an automatic RST-based discourse segmenter.

4.3 Automatic segmentation

We use an existing publicly available RST-based discourse parser for automatically segmenting texts into clauses and add several customized post-processing steps. Making use of automatically generated syntactic trees, these post-processing steps merge segments that do not contain a verb to one of their neighboring segments. For an example of such a case, see (20) above. The post-processing steps create a segmentation that more closely corresponds to our definition of situation-invoking syntactic structures, avoiding many verb-less segments which would have to be marked as *no situation*. Our aim is to make the annotation process both efficient and replicable. During manual annotation, the entire document is presented by showing one segment per line, and annotators are asked to give labels for each segment.

Automatic discourse segmentation. For automatic discourse segmentation, we use the SPADE discourse parser (Soricut and Marcu, 2003), whose discourse segmenter takes

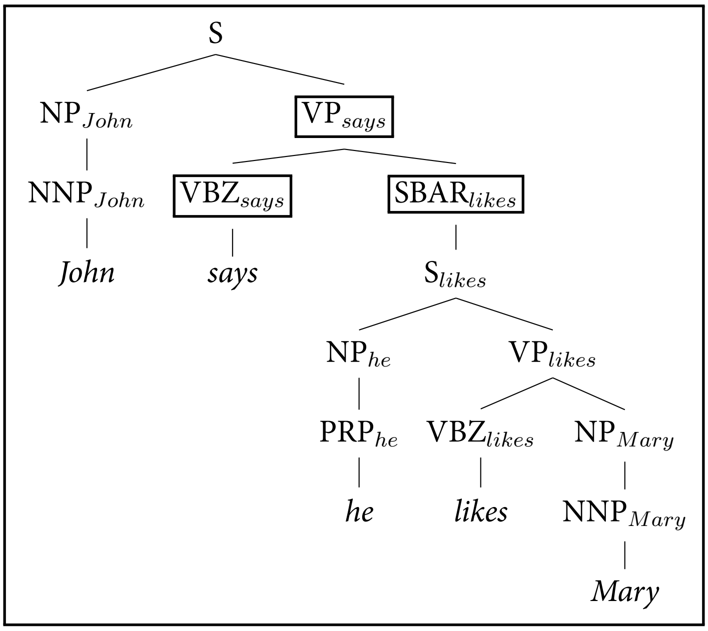

Figure 4.1: Lexicalized syntactic tree used for discourse segmentation by SPADE (Soricut and Marcu, 2003). Boxes mark the nodes used for the computation of the probability of inserting a boundary after the word "says".

a sentence as input and outputs the boundaries between the sentence's EDUs.[5] Sentence splitting is performed using the PTBTokenizer provided by the Stanford CoreNLP toolkit.[6]

SPADE's discourse segmenter is based on a probabilistic model learned from the RST Discourse Treebank (Carlson et al., 2001). For all words in the vocabulary, the probability for inserting a boundary after a word w is estimated from the treebank and depends on a lexicalized version of the corresponding sentence's syntactic tree (Magerman, 1995), incorporating the intuition that syntactic and lexical information interact in the process of identifying EDUs. Figure 4.1 gives an example of such a lexicalized tree. The probability of inserting a boundary after a word w is computed depending on the highest node in the tree lexicalized with w that has a right sibling. In the case of figure 4.1, the probability for inserting a boundary after "says" given the sentence's syntactic tree is computed as the fraction of occurrences of the lexicalized rule in the corpus for which a boundary (indicated by ↑) is present in the corpus:

$$P(boundary|says, tree) = \frac{count(VP_{says} \rightarrow VBZ_{says} \uparrow SBAR_{likes})}{count(VP_{says} \rightarrow VBZ_{says}\ SBAR_{likes})}$$

A boundary is inserted if the probability is greater than 0.5. Soricut and Marcu (2003) report that without lexicalizing the rules, this approach does not work. The probabilities are estimated based on gold standard syntactic parses from 6132 sentences from the Wall Street Journal section of the Penn Treebank (Marcus et al., 1993). Using automatic parses on the test set, Soricut and Marcu report that 82.7% of the automatically identified discourse boundaries are also present in the gold standard (precision), and that the system

[5] `http://www.isi.edu/licensed-sw/spade`

[6] `http://nlp.stanford.edu/software/tokenizer.shtml`

correctly identifies 83.5% of the boundaries in the gold standard (recall). This amounts to an F-measure of 0.83, compared to an upper bound given by human agreement of 0.98.

The segmentation provided by SPADE is sometimes more fine-grained than is appropriate for our task of identifying situation segments. In the following we describe our post-processing steps.

Heuristic post-processing. First, we identify segments that meet one of the following two requirements: (a) the segment lacks a verb, or (b) the segment contains only a *to*-infinitive construction. For each of the segments identified in the first step, we then determine whether it is possible to merge the segment with a neighboring segment, and which of the adjacent segments is best for doing so. For identifying the best neighboring segment, we first leverage sentence boundaries: if, as "for every dollar" in (27), the verbless segment occurs at the start of the sentence, we merge it with the immediately following segment. Correspondingly, a verb-less segment at the end of the sentence will be merged with the immediately preceding segment.

(27) **SPADE segmentation:** [For every dollar] [donated to Goodwill in 1998,] [we helped our "graduates" earn an estimated $102.]
Modified segmentation: [For every dollar donated to Goodwill in 1998,] [we helped our ”graduates" earn an estimated $102.]

In some cases, however, a verb-less segment occurs between two segments belonging to the same sentence. In these cases, we make use of the parse tree produced by the Charniak parser (Charniak, 2000) during SPADE's segmentation to determine whether the segment should be merged with the preceding or with the following segment. We attach the verb-less segment to the neighboring segment which is 'closest' in the parse (if all three segments are at the same level, we merge with the preceding segment). The procedure is as follows: first, we identify (a) the lowest common ancestor node of the preceding and the current segment and (b) the lowest common ancestor node of the current and the following segment. If one of these nodes is dominated by the other, the current segment is attached to the dominated node's segment.

We illustrate this procedure using example (28), which shows the segmentation produced by SPADE for the utterance "Remember what she said in my last letter?", and figure 4.2, which shows the corresponding parse. The second segment of SPADE's segmentation contains no verb. The lowest node in the parse tree dominating the first two segments of SPADE's segmentation is the `VP` node shown on the same line as *Remember*, and the lowest node in the parse tree dominating the last two segments is the `SBAR` node. As the `VP` node dominates the `SBAR` node, we attach the segment *what* to the segment that follows it, resulting in the modified segmentation also shown in (28).

(28) **SPADE segmentation:** [Remember] [what] [she said in my last letter?]
Modified segmentation: [Remember] [what she said in my last letter?]

As a final step, we merge all segments containing a verb but starting with "to" to their preceding segment, as illustrated by example (29). For this difference in RST and situation

```
(S1
  (S
    (VP (VB Remember)
      (SBAR
        (WHNP (WP what))
        (S
          (NP (PRP she))
          (VP (VBD said)
            (PP (IN in)
              (NP (PRP\$ my) (JJ last) (NN letter)))))))
     (. ?)))
```

Figure 4.2: Constituent parse for example (28) produced by the Charniak parser.

segmentation, see also the discussion of (25) above.

(29) **SPADE:** [He used it] [to preserve baby tomatoes, cucumbers, and strawberries in translucent cubes] [...].
Modified: [He used it to preserve baby tomatoes, cucumbers, and strawberries in translucent cubes] [...] `(MASC fiction, ``Captured moments'')`

Choice of automatic discourse segmenter. We decided to use SPADE rather than another automatic discourse segmenter as SPADE produces a relatively fine-grained segmentation, which, with some simple post-processing steps, resulted in a satisfactory approximation of situation segmentation (for the related inter-annotator agreement see section 6.3.1). A slightly more recent publicly available discourse segmenter, SLSeg (Tofiloski et al., 2009), which has been shown to outperform SPADE (though on a differently annotated data set), produces larger segments. SLSeg, also based on RST, considers adjunct clauses with finite or non-finite verbs and non-restrictive relative clauses as discourse units. In contrast to segmentation in the RST Treebank, however, complement clauses are not considered to be discourse units, and each segment is required to contain a verb. For instance, as illustrated by example (30), SLSeg avoids separate segments for complement clauses such as "that turtles live up to 100 years". In the case of situation segmentation, we need a separate segment for this complement clause which invokes an ABSTRACT ENTITY.

(30) **SPADE**: [When I was a kid,] [I didn't believe the fact] [that turtles live up to 100 years.]
SLSeg: [When I was a kid,] [I didn't believe the fact that turtles live up to 100 years.]

It proved more practical to apply a post-processing step merging some of the fine-grained segments produced by SPADE rather than finding a method for splitting SLSeg's output.

Our post-processing step merges segments that do not contain a verb to one of their neigboring segments, and we find the resulting segmentation to be a good approximation of our definition of situation segments.

4.4 Segmentation handling by annotators

During annotation, annotators see the entire document and mark segment by segment, where segments are separated by newlines. The automatic segmentation is intended as a guidance for annotators to identify situation-invoking constructions, which may not necessarily map to contiguous spans of text. In example (31), the subject of the second segment is identical with the one of the first clause. In both cases, the noun phrase "paraveterinary workers" is taken into account when assigning a segment's situation entity type.

(31) [Paraveterinary workers either assist vets in their work] [or may work within their own scope of practice.]

The automatic segmentation procedure described above produces segments containing exactly one situation most of the time (see section 6.3.1), but nevertheless sometimes fails to produce adequate segments. The segmentation provided by SPADE, designed for a slightly different task, is at times too fine-grained, and some of these cases are not caught by our post-processing step. Our rules were carefully developed using some test documents, but the final results on the entire corpus nevertheless contain segments which should not be considered as invoking a situation entity. For this reason, the annotation interface includes a way for annotators to both mark such segmentation errors and indicate how they should be corrected.

If a situation is split over multiple segments, we ask annotators to mark the segment containing the dominating verb with the appropriate situation entity type and features. We ask them to indicate for the respective other segment that it is *not a complete situation*, and that it *belongs to the previous (or following) segment*. For example, the last segment of example (32) does not invoke a situation, but is only a modifier of "an interesting one."

(32) [So the shift in the image of Gates has been an interesting one] [for me to watch.]

In addition, we provide for cases of discontinuous situations, which, for example, occur when a relative clause interrupts the main clause or when multiple infinitive constructions are attached to the same dominating verb as in example (33). We number the segments (example (33) shows one segment per line) and allow the annotator to indicate the number of the segment which forms a situation together with the current segment; specifically, the indicated segment should contain the situation's main verb. In example (33), the annotator marks for segment number 2 that the segment *belongs to the previous* *segment*, and for segment number 3 that it *belongs to segment number 1*.[7]

(33) (1) The universal priesthood of believers implies the right
(2) and duty of the Christian laity not to only read the Bible in the vernacular,
(3) but also to take part in the government and all the public affairs of the Church.
`(Wikipedia article on Protestantism)`

There also is the reverse case, segments containing *more than one situation*. In example (34), the automatic segmentation did not recognize the two different clauses in the third segment (main verbs marked bold below).

(34) [I think] [it almost does not matter] [what the judge **does** since **it's** clear] [that the balance of power is shifting again [...].]
`(Mini-MASC, Article247_3500.txt)`

Segments invoking more than one situation entity should be annotated with the situation entity type of the dominating verb ("does" in example (34)), or, if none of the verbs dominates the others, that of the first situation-invoking verb in the segment ("knows" in example (35)). Additionally, annotators are asked to give a comment explaining the second (or more generally, other) situation(s).

(35) [... because anyone] [who follows me on Facebook already **knows** that and because this post **concerns** Urbino, not my daily descent into emo.] (MASC blog Italy)

There are also segments not introducing a situation to the discourse at all, and which are not part of another situation. An example of such a situation is given in (36) below. Segments not invoking any situation which do not belong to another situation are marked with *no situation*.

(36) Cheers, Kara

4.5 Summary

In this work, inspired by the original statement of Smith (2003) that "the clauses of a text invoke situation entities," and following previous work by Palmer et al. (2007), we pre-segment the text into clauses prior to the annotation task. We have given a definition of a *situation segment*, which is inspired by the insight that discourse modes capture the aspectual nature of a text at the discourse-level. We have compared this definition to definitions of discourse segments in various theories of discourse structure, and found our definition to be most similar to that of EDUs in RST. We use a publicly available automatic discourse segmenter, SPADE, to pre-segment the texts, and add post-processing steps based on the syntactic parse trees in order to avoid many verb-less segments. Remaining segmentation errors are manually corrected as part of the annotation process. In the next chapter, we turn to the guidelines for manual annotation of these segments with situation entity types and the related features.

[7] The cases of marking that a segment belongs to the previous or following segment can be regarded as a special case of marking that the content of a segment belongs to another segment, and have been introduced for the annotator's convenience.

Chapter 5

Annotation scheme and guidelines

In this chapter, we give details on our *annotation scheme*, i.e., the set of labels we assign to situation segments, and the accompanying *annotation guidelines*. The latter are a set of rules and examples illustrating to annotators how to apply the annotation scheme to text. These detailed instructions can be used to train new annotators, and thus form the basis for the replicability of our annotation task.

Smith (2003) does not specify how to determine the situation entity type of a segment, but instead relies on descriptive characterizations and examples to convey an understanding of the types to her readers. Palmer et al. (2007) create a small corpus labeled with situation entity types, annotating clauses in an intuitive fashion reaching only moderate agreement (as detailed in Section 3.1.4). The findings from this earlier empirical work, as well as our own pilot annotation studies, indicate that some of the situation entity types are easier to recognize than others. Situation entity types are *covert* linguistic categories (Smith, 1997, p. 5), which are not marked morphologically, lexically or by sentence pattern in every sentence in which an element belonging to the category occurs (Whorf, 1945). In some cases, the situation entity type can be identified based on surface structure or clear linguistic indicators, but in other cases, a variety of factors needs to be interpreted. Internal temporal and aspectual properties of the verb constellation, the verb's arguments, as well as clause-level modifiers play a role.

In the present work, we develop an annotation scheme and guidelines for annotating situation entities with their type that are in line with Smith's original proposal. We introduce decisions for many difficult cases that were not covered by earlier empirical work, making sure that our annotation scheme is consistent with the descriptions of the relevant phenomena in semantic theory as surveyed in Chapter 2.

Specifically, our set of *situation entity type* labels comprises the categories Event, Report, State, Generalizing Sentence, Generic Sentence, Fact, Proposition, Resemblance, Question, and Imperative. Out of these, many instances of Event, State, Generalizing Sentence and Generic Sentence are relatively harder to recognize, as their identification often requires interpreting a combination of several syntactic and semantic cues. We therefore instruct our annotators to apply the following procedure for determining a situation entity's type. First, they are asked to identify Questions, Imperatives or Abstract Entities, because these types are relatively easily recognizable by their surface linguistic form. Second, for any situation entity not falling into one of these categories, annotators are asked to determine the values of three *situation-related features*, along which the situation entity types differ as described in the analysis presented in this chapter.

The first feature describes the fundamental lexical aspectual class of the segment's main verb (Section 5.1.2) as ***dynamic*** or ***stative*** and is relevant for distinguishing Event and State. The other two features are primarily relevant for identifying Generic Sentence and Generalizing Sentence, capturing whether the subject of the main verb is ***generic*** (Section 5.1.3), and whether the clause is ***episodic***, ***static*** or ***habitual*** (Section 5.1.4). In addition to its situation entity type, we require our annotators to assign the values of these features when labeling a segment. This allows a fine-grained analysis of inter-annotator agreement and the identification of reasons for disagreements. Gathering annotations at the level of the situation-related features additionally provides training data for these sub-tasks, which have partially been studied before (see Chapter 3), but for which no large annotated corpora are available.

For each situation entity type, we introduce the most prominent situation-related feature combinations. Based on the feature values, which are often easier to determine in isolation than the full situation entity type, annotators can then assign the situation entity type according to the rules presented in Table 5.1.

Situation entity type	**Main referent**	**Aspectual class**	**Habituality**
Event	***non-generic*** ***generic***	***dynamic***	***episodic***
State*	***non-generic***	***stative***	***static***
Generic Sentence	***generic***	***dynamic*** ***stative***	***habitual*** ***static*** or ***habitual***
Generalizing Sentence	***non-generic***	***dynamic*** ***stative***	***habitual***

Table 5.1: Situation entity types and their features. *Other feature combinations are possible for State, see *coercion* in Section 5.2.1.

Section 5.1 explains how the situation-related features are marked in our annotation scheme, and gives an overview of the corresponding guidelines. Section 5.2 introduces the annotation guidelines for the above listed set of situation entity types, and explains the correspondence between the situation entity types and the situation-related features. This chapter provides an overview of our annotation scheme and guidelines and explains the reasons behind choosing the present set of labels and guidelines. The complete version of our annotation guidelines is available in our annotation manual (Friedrich et al., 2015a).

5.1 Situation-related features

An important contribution of this work is to render the situation entity type annotation task feasible by breaking up the type identification into multiple steps for difficult cases. We ask annotators to determine the situation-related features described in this section first, and then use the determined values to identify the situation's type. This section addresses the situation-related features one at a time, forming the basis for Section 5.2, which explains the relationship between the values of the situation-related features and the situation entity types.

5.1.1 Main verb and main referent

Two parts of a clause provide important information for determining its situation entity type (Friedrich and Palmer, 2014b), a situation's *main verb* and its *main referent*. For English, the main verb is the non-auxiliary verb ranked highest in the dependency parse as illustrated by the following examples (main verbs marked bold).

(1) John does not **like** Mary.

(2) John has **kissed** Mary.

(3) John is going to **attend** the concert.

(4) John **entered** the room,
balancing a tray with tea and cookies.

The main referent is loosely defined as the main entity that the segment is about; and is thus related to the notion of the *topic* of an utterance (Smith, 2003, chap. 8.2). In English, this is usually the grammatical subject of the clause.[1] In all of the above examples, "John" is the main referent.

Cases where the main referent does not coincide with the grammatical subject in English include clauses applying the syntactic mechanism of topicalization (5), cleft constructions (6) and existential clauses (7).

(5) **That pizza**, I won't eat.

(6) It was **John** who wrote the letter.

(7) There are **two books** on the table.

We do not mark the spans of the main verb and main referent when presenting the data to our annotators, appealing to their intuition when identifying a situation's main verb

[1] In other languages, this may be more complex and requires more detailed annotation guidelines. To name one example, in a preliminary study on German, we show that the main referent may occur as an argument in dative or accusative case in certain constructions (*Es gruselt **mir** vor dir. / **Mich** friert es.*) (Mavridou et al., 2015; Friedrich et al., 2015a).

and its grammatical subject. As all of our annotators have at least some basic linguistic training, this was an easy task for them. If the main referent does not coincide with the grammatical subject of a clause, annotators indicate this during annotation by marking a checkbox labeled “not the grammatical subject”. As our annotation interface does not allow for marking spans, we do not require them to mark the actual mention of the main referent. Nevertheless, these cases receive one of the labels ***generic*** or ***non-generic*** as explained below.

Additionally, the label ***expletive*** is intended to capture expletive sentences such as (8), in which the grammatical subject “it” does not refer to an object or kind. This label was used infrequently and somewhat inconsistently by our annotators; for a discussion see Section 6.3.3.

(8) It was raining.

5.1.2 Lexical aspectual class of main verb

This situation-related feature addresses a situation entity’s *main verb*. Our annotation scheme provides two labels for the *fundamental lexical aspectual class* of a clause’s main verb: ***dynamic*** and ***stative***. Informally, ***stative*** usages of verbs express properties and states holding in time, while ***dynamic*** usages of verbs report that something happens, describing actions or processes which cause changes over time. We treat fundamental lexical aspectual class (henceforth: *aspectual class*) as a property of verbs in context, as many verbs may be used both in ***dynamic*** (9) and in ***stative*** (10) senses.

(9) She **filled** the glass with juice. (***dynamic***)

(10) Water **fills** the pool. (***stative***)

Our notion of the ***stative***/***dynamic*** distinction corresponds to Bach’s (1986) distinction between states and non-states; to states versus occurrences (events and processes) according to Mourelatos (1978); and to Vendler’s (1957) distinction between states and three event classes (activities, achievements, accomplishments). A survey of the relevant linguistic literature has been presented in Chapter 2.

The aspectual class (Siegel, 1998b; Siegel and McKeown, 2000) of a clause is a function of the main verb and a select group of complements; which complements are relevant may differ per verb. It corresponds to the aspectual class of the main verb when ignoring any aspectual markers or transformations. For example, English sentences with perfect aspect are usually considered to introduce states to the discourse (Smith, 1997; Katz, 2003), but we are interested in the aspectual class before this transformation takes place. At the clause-level, “John has kissed Mary” introduces a State, but the fundamental aspectual class of the verb “kiss” in this context is ***dynamic***. In other words, the task of identifying aspectual class can be regarded as a coarse-grained word sense disambiguation task.

To convey the notion of aspectual class to our annotators, we provide them with many examples, as well as linguistic tests for lexical aspect. For example, stative verbs disallow progressive constructions (11), while dynamic verbs allow only limited use of the simple present tense (12).

(11) *I am **owning** a house.

(12) *I **run** to class now.

None of these tests are absolute, they are only indicators, but they help when trying to determine a verb's lexical aspect. In addition, our annotation manual gives directions on how to treat several special cases. Some verbs seem to allow an interpretation as ***stative*** or as ***dynamic*** with a rather frequent habitual reading. Examples are "be called" (in the sense of "my name is") or "work at" (in the sense of "my employer is"). We treat such cases as ***stative***. Another unclear case are aspectual verbs ("start", "stop", "continue"). We treat them as the main verb of their respective situations, and mark them as ***dynamic***, with the exception of "continue" in certain contexts describing the persistence of a state (13).

(13) I wanted to leave, but John **continued** to sit on the wall. (***stative***)

The real-life duration of states and events is not an indicator for the verb's aspectual class. ***Stative*** verbs may describe properties that only hold for a very short amount of time (14), while events described using a dynamic verb may have long durations as well. Both clauses describing states and clauses describing events may or may not include start or end points. ***Dynamic*** processes (Bach, 1986) without natural endpoints as in (15) are considered as Events in our annotation scheme and should be marked as having ***dynamic*** aspectual class.

(14) For a minute, she **was** very puzzled. (***stative***)

(15) For twenty-thousand years, the earth was **cooling**. (***dynamic***)

Finally, there are cases where both a ***stative*** and a ***dynamic*** interpretation are possible (16). We allow annotators to indicate such cases as *both readings*.

(16) Your soul was made to be **filled** with God Himself. (***stative/dynamic***)
(Brown corpus, religion)

5.1.3 Genericity of main referent

With this feature, we capture whether the main referent of a situation is ***generic***, i.e., refers to a kind or to arbitrary members of the kind. Our guidelines are in accordance with the definitions of NP-level genericity of Krifka et al. (1995). Main referents that do not refer to a kind but to some particular entity are labeled as ***non-generic***.[2]

[2] This feature value was originally dubbed *specific*, and appears this way in Friedrich and Palmer (2014b). The term *specific* is used in the linguistic literature (e.g. Krifka et al., 1995) to describe NPs referring to specific individuals rather than *nonspecific* NPs, which do not refer to particular entities. Our feature value does not refer to this sense of *specific/nonspecific*. Both specific and nonspecific NPs can be ***generic*** in our scheme: "The lion is a dangerous animal" (specific, kind-referring) and "A lion roars when it is hungry" (nonspecific, non-kind-referring) are both marked ***generic*** in our data. Similarly, "Simba roared" (specific, non-kind-referring) and "A lion must be standing in the bush over there" (nonspecific, non-kind-referring, see Krifka et al. (1995, p. 15)) are marked ***non-generic*** because they do not make statements about kinds but rather about particular individuals.

Non-generic main referents are particular entities (17), particular groups of entities (18), organizations (19), particular situations (20) or particular instantiations of a concept (21).

(17) Mary likes popcorn. (***non-generic***)

(18) The students met at the cafeteria. (***non-generic***)

(19) IBM was a very popular company in the 80s. (***non-generic***)

(20) That she didn't answer her phone really upset me. (***non-generic***)

(21) Today's weather was really nice. (***non-generic***)

In English, definite NPs (22) and bare plural NPs (23) are the main kind-referring NP types (Smith, 2003, p. 73), though virtually all NP types – definites, indefinites (24) and quantified NPs (25), full NPs, pronouns and even proper names (e.g. species names such as "Elephas maximus") – can be found in ***generic*** and ***non-generic*** uses depending on their clausal context.

(22) The lion has a bushy tail. (***generic***)

(23) Dinosaurs are extinct. (***generic***)
Dogs are barking in the garden. (***non-generic***)

(24) A lion has a bushy tail. (***generic***)
A lion escaped from the zoo. (***non-generic***)

(25) Few/Most/Many/Some people like spinach. (***generic***)
Few/Most/Many/Some people ate spinach at the party yesterday. (***non-generic***)

While some NPs clearly make reference to "well-established kinds", other cases are not so clear cut. For example, in Western cultural context, "the Coke bottle" is considered to be a well-established kind, but it is less clear for "the green bottle" (Krifka et al., 1995, p. 11). Even for cases such as the latter, humans tend to make up a context in which the NP describes a kind. Sentence (26) gives an example for such a case: While "lions in captivity" are not a generally well-established kind, the NP does not describe a particular group of lions in this context.

(26) Lions in captivity have trouble producing offspring. (***generic***)

Our annotation scheme does not rely on a particular ontology of classes which would, for instance, define that "lions" are a kind but that "lions in captivity" are a group of particular instances. Instead, we instruct our annotators to decide on whether a noun phrase refers to a kind or not based on its interpretation in the respective context. Cases such as (26), which make a statement about a kind or characterize arbitrary members of a kind, are labeled as ***generic***.

Gerunds may occur as the subject in English sentences. They usually describe some process or kind of process, and thus also do not have a clear physical referent. When they describe a particular process such as in (27a), we mark them as non-generic individuals. When they describe a kind of process as in (27b), however, we mark them as concepts.

(27) (a) Knitting this scarf took me 3 days. (***non-generic***)
(b) Knitting a scarf is generally fun. (***generic***)

Finally, there are cases such as (28) or (29) where the subject NP directly refers to the kind or species rather than its members. Both cases are marked as **generic**.

(28) A new species of foxes was discovered. (***generic***)

(29) One species of foxes has grey ears. (***generic***)

In addition, the label ***cannot decide*** may be used for cases where the annotator does not feel comfortable choosing one of the other labels. For example, it is unclear whether the subject NP in (30) refers to a kind or not.

(30) The students of Saarland University don't mind eating at their mensa. (***cannot decide***)

5.1.4 Habituality of clause

While the fundamental lexical aspectual class addresses a word-sense level feature of the main verb, habituality is a situation-related feature at the clause-level. According to our annotation scheme, clauses are classified as one of the three categories ***episodic***, ***habitual*** and ***static***. In our annotation guidelines for this feature, we follow Carlson (2005). ***Episodic*** clauses report particular events (31), and ***habitual*** clauses constitute generalizations over events and activities (32), or even states (33).

(31) Mary cycled to work today. (***episodic***)

(32) Mary usually cycles to work. (***habitual***)

(33) Sloths sometimes rest on top of branches. (***habitual***)

Habitual clauses may also have ***generic*** subjects, generalizing over the members of a kind and situations at the same time (34). In the case of clauses whose subject refers to kinds, ***habitual*** clauses include cases where a situation occurs regularly for (possibly different) members of the class. For example, in (35), a single spider can die only once, but the sentence is habitual as it generalizes over situations in which spiders die. If the subject is ***non-generic***, however, the situation must repeat for the same subject, otherwise, it is ***episodic*** or ***static***.

(34) Soap is used to remove dirt. (***habitual***)

(35) Spiders die in autumn after producing an egg sac. (**habitual**, **generic** main referent, Generic Sentence)

When introducing this feature, we were primarily interested in the distinction of ***episodic*** and ***habitual*** for verbs with ***dynamic*** lexical aspectual class, but it quickly became clear that for a full-text annotation, the additional label ***static***[3] is required for clauses that are neither ***episodic*** nor ***habitual***. These are clauses with ***stative*** lexical aspectual class, and clauses stativized for other reasons including negation, modality and perfect aspect. English clauses in past or present perfect such as (36) are ***static***, as they focus on the post-state of an event rather than the event itself (Katz, 2003).

(36) Mary has made a cake. (***static***)

Modalized (37) and negated sentences (38) tend to be ***static***: they do not express information about a particular event, but refer to actual or possible states of the world.

(37) Mary can swim. (***static***)

(38) Mary didn't go swimming yesterday. (***static***)

The above definitions of habituality and stativity are generally agreed upon in literature. However, the interaction of these phenomena is by no means trivial (Hacquard, 2009), and required making some decisions during the design of our annotation guidelines. Here, we explain these decisions, which are all motivated by a clause's entailment properties.

One difficult issue is how to interpret and mark negated sentences such as (39a) whose positive version (39b) is habitual.

(39) (a) John does not smoke. (***habitual***)
(b) John smokes. (***habitual***)

Sentence (39a) can be considered either ***static*** because of the negation (*It is not the case that John smokes*), or as ***habitual*** because it characterizes John's behavior (*In any relevant situation, John does not smoke*). Both decisions are possible (Garrett, 1998), and we decide for the latter possibility. This decision is supported by the observation that (39a) is similar in its entailment properties to (40), which due to the frequency adverbial "never" clearly generalizes over relevant situations (though note that this is not a linguistic test).

(40) John never smokes. (***habitual***)

Likewise, we mark sentences containing modal verbs as habitual if they have a strong implicature that an event has actually happened regularly (Hacquard, 2009), as in (41). In contrast, (37) is static as it does not imply that Mary actually swims regularly.

(41) I had to eat an apple every day. (***habitual***)

[3] For clarity, we use the label ***static*** for the clausal aspect of non-episodic and non-habitual sentences. We reserve ***stative***, which is more common in the literature, for the lexical aspectual class.

5.2 Situation entity types

We now describe the annotation guidelines for the set of situation entity types in our annotation scheme, making reference to the situation-related features explained in the previous section. Following Smith (2003) and Palmer et al. (2007), we group these types as sub-types of Eventuality, GENERALIZING SENTENCE and SPEECH MODE.[4]

5.2.1 Eventualities

STATES (42) and EVENTS (43) are subsumed under the type Eventuality. According to Smith (2003), "events take place in time,...; states are specific situations that hold in time." The most important difference between STATE and EVENT is the ***stative-dynamic*** distinction of their lexical aspectual class. EVENTS must in general have main verbs with ***dynamic*** lexical aspectual class, while ***stative*** main verbs indicate STATES. In addition, all EVENTS are ***episodic***, expressing that something happened or is happening, while all STATES are ***static***.

(42) John knows the answer. (STATE, ***stative***, ***static***)

(43) Mickey painted the house. (EVENT, ***dynamic***, ***episodic***)

When distinguishing STATES from EVENTS, a frequent difficult case are past participles that are used as reduced relative clauses as in (44). In reading (a) the participle refers to an event, and the reduced relative clause is a passive construction. In reading (b) the participle functions as an adjective and hence describes a property of the noun it post-modifies. We instruct annotators to decide on a case-to-case basis which reading they find to be more prominent and annotate the situation entity accordingly.

(44) A movie, filmed in black an white, ...
(a) A movie, which was filmed in black and white, ... (EVENT, ***dynamic***)
(b) A movie, which is filmed in black and white, ... (STATE, ***stative***)

Clauses whose main verb has ***dynamic*** lexical aspectual class can be shifted to the situation entity type STATE by the mechanism of *coercion* (Moens and Steedman, 1988; Smith, 1995). We consider negation, modals, future tense and perfect aspect to shift the situation type, as well as statements occurring as conditionals. Clauses with progressive ***dynamic*** main verbs are still considered as (ongoing) EVENTS.

(45) Mickey was painting the house. (EVENT, ***dynamic***, ***episodic***)

Negated (46) or future (47) events and expressions within the scope of a modal verb (48)

[4] The latter supertype was called SPEECH ACT in earlier publications. This was a possibly confusing choice as the distinction between QUESTION and IMPERATIVE on the one hand and the other situation entity types on the other hand is not one in Searle's (1969) sense of speech acts, but is rather related to sentence mode.

do not state that an event takes place in fact, but introduce the possibility of an event. The situation entity type is coerced to STATE in such cases.

(46) John did not win the lottery. (***dynamic***, STATE)

(47) John will move to his own place next week. (***dynamic***, STATE)

(48) She might have left. (***dynamic***, STATE)

As explained above, the English perfect stativizes clauses, focusing on the circumstances of an action's being completed at the time of reference. Therefore, clauses with a ***dynamic*** main verb in perfect aspect are marked as STATE (49).[5]

(49) She has left. / She had left. (***dynamic***, STATE)

In addition, adverbials such as *probably, likely, certainly* can transform a situation entity into a STATE (50).

(50) John probably argued with his parents yesterday. (STATE)

We also consider conditional clauses to introduce possible events or states, and mark them as STATES (51).

(51) If John had won the lottery, (***dynamic***, STATE)
he would not be living with his parents. (***stative***, STATE)

EVENTS may also have a ***generic*** main referent if they describe a particular happening related to the kind (52). In contrast, STATES have ***non-generic*** main referents, with the rare exception of EVENTS relating to kinds which are coerced to a STATE on the clause level as in (53).

(52) In September 2013 the blobfish was voted the "World's Ugliest Animal". (EVENT, ***generic***, ***dynamic***, ***episodic***) `(Wikipedia)`

(53) The wheel has been invented. (STATE, ***generic***, ***dynamic***, ***static***)

[5] This is different for the German perfect. In German, some clauses in perfect clearly refer to an EVENT ("Wir sind gestern ins Kino gegangen" – "We went to the movies yesterday"), while other clauses focus on the post-state of an action ("Ich habe schon gegessen" – "I have eaten"). Finally, there are underspecified cases such as "Sie haben mir den Job gegeben" ("They gave / have given me the job"), for which we introduced an additional label EVENT-PERF-STATE in a preliminary study (Mavridou et al., 2015).

5.2.2 General Statives

General Statives do not express particular states or events. Instead, they express regularities of events or properties of kinds of entities, and thus differ from the above Eventualities in their entailment properties and aspectual nature. General Statives have two subtypes: Generic Sentences and Generalizing Sentences.

Generic Sentences are defined as clauses making a generalizing statement about a kind or class or arbitrary members of the kind, i.e., clauses having a ***generic*** main referent according to the above definition. They have either a ***stative*** main verb (54) or a ***dynamic*** main verb with a ***habitual*** reading (55).

(54) Computers are very useful. (Generalizing Sentence, ***generic***, ***stative***, ***static***)

(55) Lions eat meat. (Generalizing Sentence, ***generic***, ***dynamic***, ***habitual***)

In our definition, Generalizing Sentences are clauses reporting regularities about ***non-generic*** main referents (56). They have ***habitual*** main verbs in simple present or simple past tense, implying that some event happens regularly (56) or that some state is taken on repeatedly (58).

(56) John drives to work. (Generalizing Sentence, ***dynamic***, ***habitual***)

(57) John always fed the cats last year. (Generalizing Sentence, ***dynamic***, ***habitual***)

(58) I often feel as if I only get half the story. (Generalizing Sentence, ***stative***, ***habitual***)

Our definition of the two subcategories of General Statives is somewhat simplified from that by Smith (2003). We consider as Generic Sentences *all* clauses that refer to something typically holding of a class or kind. We make no distinction between whether the clause states inherent properties of the kind, describes actions carried out repeatedly by the kind, or describes something that is usually done with instances of the kind. It is the latter case on which we differ from previous work. According to Smith (2003, p. 73), kind-referring NPs may also occur in Generalizing Sentences, as in (59).

(59) Potatoes are served whole or mashed as a cooked vegetable. (Generic Sentence)[6]

While we can see that this example describes a pattern of what people usually do with potatoes rather than an inherent property of potatoes, we prefer the more unified notion of Generic Sentences described above.

The distinction between Generic and Generalizing Sentences is far from being a clear one, and researchers have developed many different notions of *genericity* (Carlson, 1995). The common denominator of Generic Sentences and Generalizing Sentences is that at least one of the two basic varieties of genericity applies (Krifka et al., 1995): (a) the

[6] This particular example is classified as a Generalizing Sentence by Smith (2003, p. 73).

clause generalizes over members of a kind or directly makes a statement about a kind; or (b) the clause generalizes over situations. We allow annotators to choose the supertype General Stative if they are not sure whether a clause should be labeled with GENERIC SENTENCE or GENERALIZING SENTENCE, as these two situation entity types have a similar function when determining a passage's discourse mode. However, as we will report in the following chapter, most often, disagreements occur between STATE and GENERIC SENTENCE rather than between the two subtypes of General Stative.

In contrast to EVENTS, General Statives are not subject to coercion to other situation entity types when the clause is negated, modalized or in future tense, as illustrated by (60) and (61). General Statives are already stative in nature. We label situation entities as GENERIC SENTENCE or GENERALIZING SENTENCE rather than STATE to capture generalizations over kinds or situations. This definition also applies to negated, modalized or future-tense clauses.

(60) Whales will not be extinct in 100 years. (GENERIC SENTENCE)

(61) John will drive to work from next week on. (GENERALIZING SENTENCE: a repeated action in the future)

Modification by modals expressing belief does not shift the situation type of General Statives (62). However, modals like *may, can, could* and *must* express ability, possibility or necessity rather than the fact that something is done regularly, and are marked as STATE (63).

(62) Mickey probably paints houses. (GENERALIZING SENTENCE, ***habitual***)

(63) Mickey can/may/could/must paint houses for the rest of his life. (STATE, ***static***)

Modalized GENERIC SENTENCES are marked as ***static***, but they are not coerced to another situation entity type, as shown in (64).

(64) Kangaroos can/may/could/must jump all the time. (GENERIC SENTENCE, ***static***)

5.2.3 Abstract Entities

The situation entity types subsumed under ABSTRACT ENTITIES comprise FACT and PROPOSITION. We use these labels in a very particular sense here to refer to a small number of linguistic constructions which serve a particular discourse function. By referring to an event, rather than directly describing its occurrence, an author introduces the event as either something known to be true (FACT) or believed to be true (PROPOSITION).

(65) I know (STATE, licensing predicate)
that Mary refused the offer. (FACT, EVENT)

(66) It was unlikely (STATE, licensing predicate)
that Mary would refuse the offer. (PROPOSITION, STATE)

According to Smith (2003), facts are not in the world, they are about the world. While EVENT-type situation entities (67) have the effect of advancing narrative time, there is no such advancement when the same event is referred to in a factive or propositional construction.

(67) Mary refused the offer. (EVENT)

In this annotation project, we label only situation entities that appear as clausal complements of certain predicates as ABSTRACT ENTITIES. For instance, the clausal complement of *know* in Example (65) refers to a FACT, and the clausal complement of (66) refers to a PROPOSITION.

Asher (1993) defines *abstract objects* to include propositions, properties, states of affairs and facts, which, in contrast to *eventualities* (states and events), have no spatio-temporal location. In Asher's account, abstract entities can be referred to by sentential nominals (see Section 4.1) and predicate nominals (common noun phrases and verb phrases). Our definition of ABSTRACT ENTITIES is thus narrower than Asher's, the situation entities labeled as ABSTRACT ENTITIES are a subset of his abstract objects. Smith (2003, sec. 4.3) notes that there are also sentences directly expressing FACTS and PROPOSITIONS, but these cannot be distinguished on linguistic grounds. We follow her approach, not addressing these within our guidelines for situation entity type annotation.

Multiple type annotation. The clausal complement introducing an ABSTRACT ENTITY itself has a situation entity type. Annotators are asked to additionally mark the situation entity type of the embedded clause. For example, in (65) the clausal complement of "know" describes an EVENT. This procedure has the advantage that the embedded situation entities can be used as additional training material for the other situation entity types.

5.2.4 Speech Mode types

English has four types of *sentence modes*: *declarative* and *conditional* sentences make statements about the world or introduce dependencies between circumstances or events, and are marked according to one of the above explained situation entity types. Sentences in the *interrogative* or *imperative* mode, however, do not fit in any of those categories. Therefore, Palmer et al. (2007) introduce QUESTION and IMPERATIVE as two additional situation entity types to allow for an exhaustive annotation of arbitrary text.[7]

QUESTIONS (68) and IMPERATIVES (69) in English are usually easily identifiable by their surface form. Questions can also be posed indirectly or embedded in reported speech

[7] Palmer et al. (2007) and Friedrich and Palmer (2014b) call them SPEECH ACTS, as they are clauses which are purely performative. However, the term *speech act* evokes a broader definition according to Austin (1975) and Searle (1969); by uttering sentences of *any* mode the speaker may both convey information and perform a speech act. We therefore rename this situation entity type to SPEECH MODE.

(70).

(68) Wouldn't John be a good teacher? (Question)
Do you think John would be a good teacher? (Question)

(69) Stay calm. (Imperative)
Don't worry about it. (Imperative)

(70) He asked me (Event)
whether I would like to sing. (Question)

Annotators were also allowed to directly chose the label Speech Mode. The only case where we observed this in our corpus was for speech acts of thanking, e.g., "thank you."

Chapter 6

Corpus data and agreement

This chapter gives a detailed overview of our text corpus annotated for situation entity types and discourse modes. We describe the corpora chosen for this annotation project, which include texts from MASC and Wikipedia. The creation of a text corpus manually labeled with linguistic categories requires estimating to what extent the annotated data are reliable. Artstein and Poesio (2008) give a comprehensive survey of existing methods for measuring agreement in computational linguistics. In the construction of our corpus, we followed their suggestions for best practices and measuring agreement. We here give statistics on the inter-annotator agreement (reliability) and intra-annotator agreement (stability) and report on the gold standard labels created from the annotations. Finally, based on the observations we make during the agreement study, we give a critical analysis of the annotation guidelines and scheme by discussing several hard cases, and suggest directions for further refinement of the annotation guidelines.

6.1 Corpus data

Our corpus data is drawn from two resources, MASC and Wikipedia. This section describes the contents and sizes of each collection of texts.

6.1.1 MASC data

The Manually Annotated SubCorpus (MASC) is a collection of contemporary American English (Ide et al., 2008, 2010).[1] It consists of written and spoken data, a subset of roughly 500,000 words from a number of different genres drawn from the Open American National Corpus (Ide and Macleod, 2001; Ide and Suderman, 2004). For our purpose, we concentrate on the written part of MASC.

MASC is ideal for our annotation endeavor precisely because it contains texts from a variety of genres, as well as further manual annotations such as part-of-speech information, lemmas, and Penn TreeBank-style syntactic trees. Additional manual annotation layers,

[1] MASC is freely available from `http://www.anc.org/data/masc`.

which we expect to be highly useful for further linguistic investigations related to our own annotation layers, such as coreference information, discourse markers and gold standard clause boundaries, are underway at the time of this writing.

We annotate 12 out of the 20 MASC subsections, each corresponding to a genre as listed in Table 6.1. In the following, we briefly give an overview of the genres' contents. The email data contains emails from the ENRON corpus (Klimt and Yang, 2004) and from public mailing lists discussing technical issues. While the blog posts are also in an essay style, the texts of the essay genre are more formal, have been edited and published. The blog documents also include comments from readers. Each of the joke documents groups multiple jokes. The fiction genre contains seven comparably longer novels, while the ficlets genre consists of 152 short literary texts. The government documents (gov't documents) contain financial, political and military reports. The corpus also contains fund-raising letters and travel guides. From the technical genre, which contains technical papers, we omit some lines of text (e.g. programming code).

Table 6.1 shows the number of documents (as provided by MASC) and tokens. The number of tokens is obtained by processing the texts with the Stanford PTBTokenizer (Manning et al., 2014). The number of segments is the number of automatically created situation segments (see Chapter 4). The column titled "situation entities" lists the number of situation segments that received a situation entity type label in the gold standard (see Section 6.6).

Genre	Documents	Tokens	Segments	Situation entities
blog	21	33146	3592	3144
email	80	33724	3978	2252
essays	7	26490	2380	1978
ficlets	5	33864	5452	3963
fiction	6	38847	4820	4651
gov't documents	5	27737	2246	2012
jokes	16	32420	4184	3630
journal	9	24386	2264	2066
letters	49	26218	2421	1981
news	52	30664	2823	1185
technical	8	22326	1669	1412
travel	7	27256	2196	2059
total	265	357078	38025	30333

Table 6.1: MASC subsections (genres) annotated for situation entities and discourse modes.

6.1.2 Wikipedia data

One genre missing from the variety of genres offered by MASC is encyclopedic data. Especially for studying the linguistic phenomenon of genericity, we were interested in this genre, and collected a set of 102 texts from the English part of the free online encyclopedia Wikipedia.[2] We use about 70 sentences from each article, starting at the beginning of each article and respecting existing paragraph boundaries. We manually assign a category label to each document (Friedrich et al., 2015b), see Table 6.2.

We chose these categories because we aimed at creating a corpus that is balanced in the sense that it contains many generic and non-generic sentences, and also generics from many different domains. For example, some sentences make statements about a "natural" kind. The subject NP in example (1) refers to a biological species, which is, due to common agreement, a "well-established kind" (Krifka et al., 1995, p.11).

(1) Blobfish are typically shorter than 30 cm.

Indefinite singular NPs, such as the subject in (2), do not directly refer to kinds and get their generic reading only when occurring in a characterizing sentence (Krifka et al., 1995). In our corpus, cases such as (2) are also marked as ***generic***, as the subject NP refers to an arbitrary clavinet here.

(2) A clavinet played through an instrument amplifier with guitar effect pedals is often associated with funky, disco-infused 1970s rock.

Other sentences in this corpus part express definitions such as the rules of a football game (3). The subject NP does not refer to a particular team here, but to any "offensive team" playing the game.

(3) The offensive team must line up in a legal formation before they can snap the ball.

6.2 Annotation procedure and development of guidelines

One typical difficulty arising in annotation efforts in the field of computational linguistics is the need to revise the annotation scheme or annotation guidelines. Recall from Chapter 5 that *annotation scheme* refers to the set of labels, categories and their possible values, and *annotation guidelines* refers to the specification of how to assign those categories to particular examples. While our annotation scheme stayed more or less stable across time, we undertook several changes to the annotation guidelines.

We started with an annotation scheme inspired by the descriptions and examples of Smith (2003), and an unpublished annotation manual by Shore and Palmer (2011). Four annotators, paid students of computational linguistics and English, applied the scheme and guidelines to several Wikipedia documents not included in the final corpus and to the written portion of Mini-MASC,[3] a selection of 2785 tokens from MASC from four different genres. Analysis of their disagreements and their feedback led to revision of the guidelines and a first stable version of the annotation manual (Friedrich and Palmer, 2014b). In

[2] `http://en.wikipedia.org`

[3] Mini-MASC is available at `http://www.anc.org/MASC`.

Category	Articles	Tokens	Segments	Situation entities
animals	13	26765	2223	1991
biographies	7	7681	641	568
botany	6	7501	655	603
organized crime	4	3932	306	277
ethnic groups	8	11802	993	889
games	5	8937	684	618
medicine	7	6046	482	414
music	12	12982	1083	898
politics	16	20627	1573	1397
religion	8	12455	974	844
science	8	11453	950	840
sports	8	17859	1420	1268
sum	102	148040	11984	10607

Table 6.2: Wikipedia data annotated for situation entities and discourse modes.

a second iteration, the guidelines were refined with regard to the genericity of the main referent (Friedrich et al., 2015b), strictly following the definitions of kind-reference by Krifka et al. (1995). A final, finer-grained adjustment was made to the guidelines related to the interaction of habituals and modality (Friedrich and Pinkal, 2015b).

In the final corpus as presented in this chapter,[4] annotations from annotators only involved in the early stage of the project were removed and the data has been re-annotated by annotators joining at a later stage, i.e., annotators who are only familiar with the final version of the manual. For some small changes in the guidelines, all annotated data were reviewed and labels were adapted by the respective annotators themselves. Each of the annotators joining later on was given a short training on Wikipedia documents not included in the corpus. Annotators were not allowed to communicate with each other about questions regarding the annotation task. In some cases, when agreement was very low, annotators were asked to review those documents, without seeing each other's annotations.

Each of the corpus sections has been annotated by three annotators. The corpus sections were assigned to annotators depending on their availability. An anonymized listing of which annotator marked which corpus section is given in Appendix A, and the anonymized identity of the respective annotators is also included in the final published corpus.

In the following, we first take a look at *inter-annotator agreement*, i.e., how often different annotators agree or disagree on the same annotation task, and then give results on an *intra-annotator agreement* experiment, in which we estimate to what degree annotators are able to reproduce their own judgments.

[4] Freely available from `http://www.coli.uni-saarland.de/projects/sitent`

6.3 Inter-annotator agreement

In this section, we give a detailed analysis of inter-annotator agreement with the aim of determining which parts of the annotation scheme result in reliable, i.e., reproducible, data, and which parts are difficult to apply for annotators. We start by comparing the various annotators' decisions of when or when not to label an automatically created situation segment with a situation entity type. We then measure inter-annotator agreement for situation entity types and the situation-related features, and analyze the respective labels one at a time with regard to how easily they are distinguished from other categories. In addition, we take a look at which labels are most often confused with which other categories.

6.3.1 Agreement on correcting situation segmentation

As explained in Section 4.3, we split the texts into situation segments automatically. Annotators are given the option to indicate a segmentation error and refrain from giving a situation entity type label in such cases. Our first question in analysing agreement is hence how often annotators agree on whether a segment invokes a situation entity or not.

Table 6.3 shows the percentage of automatically created situation segments marked as invoking a situation by either all three, two, just one or none of the annotators who labeled the respective genre. The cases marked as situation-invoking by all three annotators and the cases marked as non-situation-invoking by all three annotators constitute the cases of perfect agreement. For a total of 7.0% of all situation segments in the MASC data and in the 5.6% of all situation segments in the Wikipedia data, the set of three annotators did not unanimously agree on whether the segment invokes a situation or not. Cases that received a situation entity label by at least two annotators are added to the gold standard, i.e., adding up the percentages of the first two columns in Table 6.3 corresponds to the percentage of segments presented as invoking a situation entity in Tables 6.2 and 6.1.

The performance of our automatic situation segmentation method differs across MASC genres and Wikipedia categories. The high number of non-situation-invoking segments in ficlets and email is due to lines containing header information about the author or sender, email addresses or subject.

There are also cases where our automatic segmentation method fails to separate situations. Table 6.4 gives the percentage of automatically created segments labeled as containing multiple situations by either all three, two or one annotator(s). For such cases, annotators are instructed to annotate the situation entity included in the segment whose main verb would appear highest in the dependency tree, and the first situation entity if all main verbs rank equally according to this criterion.

	Annotators			
Genre	**3**	**2**	**1**	**0**
blog	82.2	6.1	2.3	9.4
email	53.4	3.1	5.6	37.8
essays	83.2	3.9	2.5	10.4
ficlets	66.6	2.3	2.7	28.5
fiction	90.6	4.0	1.6	3.8
govt-docs	87.6	2.6	4.0	5.8
jokes	81.6	5.1	4.9	8.3
journal	87.2	4.2	3.6	5.0
letters	77.4	4.4	5.2	13.0
news	89.2	4.0	2.2	4.5
technical	84.5	1.7	2.7	11.1
travel	90.2	3.0	1.5	5.2
all	79.4	3.8	3.2	13.6

(a) MASC, by genre.

	Annotators			
Category	**3**	**2**	**1**	**0**
animals	86.6	3.0	2.0	8.4
biographies	87.4	1.2	3.1	8.3
botany	90.2	1.8	2.1	5.8
crime	88.9	1.6	2.9	6.5
ethnic groups	87.2	2.3	3.0	7.5
games	87.3	3.1	2.3	7.3
medicine	80.9	5.0	4.6	9.5
music	80.7	2.2	3.1	13.9
politics	86.6	2.2	2.0	9.2
religion	82.6	4.0	3.6	9.8
science	85.9	2.5	4.7	6.8
sports	86.1	3.2	3.2	7.5
all	85.8	2.7	2.9	8.6

(b) Wikipedia, by category.

Table 6.3: Observed agreement on whether an automatically created situation segment invokes a situation: percentage of segments marked as invoking a situation entity by 3, 2, 1 or none of the annotators.

Corpus	**Automatically**	**Annotators**		
	created segments	**3**	**2**	**1**
MASC	38025	1.3	2.9	3.9
Wikipedia	11984	1.1	3.4	3.9

Table 6.4: Observed agreement on whether an automatically created situation invokes multiple situations: percentage of segments marking as invoking multiple situation entities by 3, 2 or 1 annotator.

6.3.2 Agreement on situation entity types

In this section, we describe and analyze the inter-annotator agreement on the assignment of situation entity types. We compute agreement over the types State, Event, Report, Generic Sentence, Generalizing Sentence, Imperative and Question, parallel to the categories used in our automatic classification experiments in Section 8.5. Abstract Entities will be discussed in Section 6.3.5.

Agreement measure. We measure agreement using Fleiss' κ (Fleiss, 1971), a generalization to multiple annotators of the agreement measure Cohen's κ (Cohen, 1968). Fleiss' κ is computed as

$$\kappa = \frac{observed\ agreement - expected\ agreement}{1 - expected\ agreement}$$

where the *observed agreement* is the proportion of pairwise judgments that match. The *expected agreement* is the proportion of pairwise judgments that are expected to match if one assumes that annotators assign the categories randomly according to some underlying distribution. Here, this distribution is estimated by assuming the same prior distribution of labels by all annotators as computed from the sample set. The expected agreement is computed as follows, with i being the number of items, and c the number of annotators (coders). The number of categories is given by k, and n_k indicates how often category n_k was assigned to an item in the data (see also Artstein and Poesio, 2008).

$$expected\ agreement = \frac{1}{(ic)^2} \sum_k (n_k)^2$$

We compute Fleiss' κ for all cases that are part of the gold standard, i.e., for all cases where at least two annotators gave a situation entity type label. Because Fleiss' κ is computed as agreement on pairwise judgments, we can compute this statistic for the subset of pairwise judgments where both annotators gave a situation entity type label. We use a modified version of the DKPro Agreement (Meyer et al., 2014) for doing so. This way of computing agreement excludes disagreements on the decision of whether a segment invokes a situation and focuses on agreement on identifying the situation entity type of a segment. Agreement on the segmentation decision is reported above in section 6.3.1.

Analysis. Aggregating pairwise judgments over the different annotators per genre, we find that observed agreement for situation entity types on the MASC data is 79.2%, and that observed agreement for the Wikipedia data is 78.2%. Expected agreement is 33.4% and 35.6% for the MASC and Wikipedia data respectively. Overall agreement amounts to $\kappa = 0.69$ for MASC and for Wikipedia to $\kappa = 0.66$. These numbers are in the range of substantial agreement according to Landis and Koch (1977), indicating good quality of our annotated corpus, but they also show that the task of situation entity annotation is far from being trivial for humans. In order to shed some light on which categories our annotators had most difficulties with, we apply Krippendorff's diagnostics (Krippendorff,

1980) for category distinctions. For each category, we compute Fleiss' κ for an artificial set-up in which all categories except one are collapsed into an artificial OTHER category. A high value indicates that annotators can distinguish this category well from others.

Table 6.5 shows the outcome of this analysis along with the relative frequencies ($n_k/(ic)$) of the situation entity types as assigned to items by the annotators. If a category is infrequent, it is harder to obtain a high κ-score (Di Eugenio and Glass, 2004), so κ-like agreement figures can only be interpreted in relation to this underlying distribution.

Situation entity type	**Frequency $n_k/(ic)$**		**Fleiss' κ**	
	MASC	**Wikipedia**	**MASC**	**Wikipedia**
all types	-	-	0.69	0.66
STATE	49.8	24.7	0.68	0.59
EVENT	25.1	19.9	0.74	0.73
REPORT	4.2	0.6	0.83	0.28
GENERIC SENTENCE	8.2	49.7	0.45	0.72
GENERALIZING SENT.	4.7	3.8	0.46	0.35
QUESTION	3.2	0.1	0.91	0.90
IMPERATIVE	3.2	0.2	0.93	0.95
NONE	1.6	1.0	n/a	n/a

Table 6.5: Situation entity types: frequencies of labels assigned and Krippendorff's diagnostics.

Table 6.5 shows that QUESTION and IMPERATIVE are the situation entity types that are easiest to identify, both obtaining high κ-scores in both corpus parts despite their low frequency. Identifying them is not trivial as they are sometimes embedded in other sentence constructions, e.g., as reported speech, and annotators do not always agree in these cases. STATE and EVENT are both frequent types and agreement is substantial. The low score of REPORT in the Wikipedia data is due to its low frequency in this corpus part; in MASC, which contains a higher percentage of segments labeled as REPORT, agreement on identifying this category is very high. Agreement on GENERIC SENTENCE is low for MASC, but it is hard to interpret as this type is infrequent in MASC. One aim of collecting the Wikipedia data was to determine whether a higher κ-score would be reached if the data set contained more items of this category; this is indeed the case. GENERALIZING SENTENCES, habitual clauses with a ***non-generic*** main referent, are the remaining difficult case: none of the two corpus parts contains a sufficient number of situation entities labeled with this type to allow for reliable agreement interpretation. However, as we will report below, agreement on identifying ***habitual*** clauses is in general good. Note that the percentage of NONE cases is only relative to the number of situation segments included in the gold standard; i.e., situation segments receiving no situation entity type label by two or more annotators are not included in this statistic.

Table 6.5 helps to identify "easier" and "difficult" categories. It does not, however, show which categories were confused with each other. In a two-annotator setting, this question is usually answered by evaluating a confusion matrix. As we have three labels per item, this matrix would be three-dimensional. To analyze label confusion we instead create a normalized version of a *coincidence matrix*. A coincidence matrix (Krippendorff, 1980, p. 149) records for each pair of labels how often an item was assigned this combination of labels by two annotators, summing over the pairwise assignments by the entire set of annotators. Each pairing of labels is entered into the table twice, using each annotator's label as the row and as the column indicator once. This coincidence matrix is also the basis for computing Fleiss' κ; the observed agreement is the sum of the cells on the diagonal divided by the total number of pairings.

Table 6.6 and Table 6.7 show the coincidence matrices for situation entity types as labeled by pairs of annotators in the MASC and Wikipedia data, respectively. In order to make the matrices readable, we have normalized the matrices per row. The data in these normalized matrices now lets us answer the question of which pairings of labels were often assigned to the same situation segment.

In MASC, many items labeled Generic Sentence received the label State by the respective other annotator. In both MASC and Wikipedia, Generalizing Sentence, which has the lowest κ-score in both corpus parts, has been confused mostly with State and Generic Sentence. In Wikipedia, we observe that many Reports are labeled as Event, but the total number of Report is very low in this genre. We manually inspected cases that are labeled as Report by at least one annotator, and with something other than Event and Report by at least one annotator. We found that they are all cases reporting on someone's position or attitude towards something. They mark situation entities with main verbs like *argue, state, say, define, claim, suggest, indicate, conclude* and *postulate*. Though Report is intended for these cases of attribution as well, our guidelines did not point this out clearly enough, such that some annotators got the impression that Report was only for situation entities introducing direct or indirect speech.

Overall, the situation entity type State is the one that is most confused with other categories. State captures not only lexically stative clauses but also a large number of coerced clauses, and the high confusion with Generic Sentence is due to the difficulty of determining the main referent's genericity in many cases, which we will analyse next.

Situation entity type	State	Event	Report	Generic	Generalizing	Question	Imperative	None	Freq.
State	**83.2**	6.6	0.2	5.9	2.2	0.3	0.2	1.4	49.8%
Event	13.0	**79.3**	2.1	1.4	2.5	0.2	0.1	1.4	25.1%
Report	2.4	12.6	**83.6**	0.3	0.5	0.1	0.3	0.3	4.2%
Generic Sentence	36.1	4.3	0.2	**48.7**	8.1	0.4	0.3	1.8	8.2%
Generalizing Sent.	23.2	13.2	0.4	14.0	**47.3**	0.2	0.1	1.6	4.7%
Question	5.1	1.3	0.1	1.1	0.2	**84.9**	0.6	6.7	3.2%
Imperative	3.7	0.6	0.4	0.9	0.2	0.6	**90.9**	2.8	3.2%

Table 6.6: Coincidence matrix for situation entity types: pairings of labels by three annotators, normalized per row. MASC data.

Situation entity type	State	Event	Report	Generic	Generalizing	Question	Imperative	None	Freq.
State	**68.0**	8.9	0.5	18.0	3.0	0.0	0.0	1.5	24.7%
Event	11.0	**78.1**	1.3	6.2	2.9	0.0	0.0	0.4	19.9%
Report	23.0	44.4	**28.3**	4.0	0.3	0.0	0.0	0.0	0.6%
Generic Sentence	8.9	2.5	0.0	**85.4**	2.1	0.0	0.0	1.1	49.7%
Generalizing Sent.	19.5	15.1	0.0	27.4	**37.4**	0.0	0.0	0.5	3.8%
Question	5.0	0.0	0.0	5.0	0.0	**85.0**	0.0	5.0	0.1%
Imperative	1.5	2.3	0.0	1.5	0.0	0.0	**84.8**	9.8	0.2%

Table 6.7: Coincidence matrix for situation entity types: pairings of labels by three annotators, normalized per row. Wikipedia data.

6.3.3 Agreement on situation-related features.

For each of our situation-related features (see Section 5.1), we compute agreement across the labels listed in Table 6.8 over all situation segments that received a situation entity type label in our gold standard. We assume here that in principle, a label can be given for each of these segments, and use the label ***cannot decide*** for missing values or if the annotator has given this label explicitly. We ask annotators to always assign feature values if possible, even if they cannot decide on the final situation entity type. Nevertheless, the "missing values" may contain cases where an annotator simply neglected to give a label, but would have decided on one of the categories other than ***cannot decide***.

The highly frequent verbs "have" and "be" are almost always ***stative***. We therefore also compute agreement for the lexical aspectual class on the subset of data where, using automatically created parses, we could identify a main verb other than "have" or "be".

Table 6.9 contains the agreement scores in terms of observed agreement, expected agreement and Fleiss' κ. Agreement is substantial with the exception of the judgments on the main referent's genericity on MASC, and lexical aspectual class on Wikipedia when examining the subset of verbs other than "have" or "be". As explained earlier, the ***generic*** category occurs rarely in MASC, leading to a high expected agreement, which in turn makes it hard to obtain a high κ-score. On the more balanced Wikipedia data, agreement for genericity is also substantial.

Main referent	Aspectual class	Habituality
non-generic	***dynamic***	***episodic***
generic	***stative***	***habitual***
expletive	***cannot decide***	***static***
cannot decide		***cannot decide***

Table 6.8: Labels for situation entity types and situation-related features used for computing agreement.

	MASC			Wikipedia		
	observed	**expected**	κ	**observed**	**expected**	κ
Aspectual class	82.7%	43.4%	0.69	81.5%	48.5%	0.64
- without *have/be*	80.9%	49.1%	0.62	78.0%	52.4%	0.54
habituality	83.3%	46.2%	0.69	79.7%	41.8%	0.65
Main referent	86.4%	70.1%	0.55	83.1%	49.1%	0.67

Table 6.9: Inter-annotator agreement for situation-related features, Fleiss' κ.

We next create coincidence matrices for each situation-related feature. Table 6.10 shows the coincidence matrix for the fundamental aspectual class of the segment's main verb. Out of the segments labeled ***dynamic*** or ***stative*** by one annotator, the other annotator

	dynamic	***stative***	***cannot decide***	**Frequency**
dynamic	**84.7**	13.0	2.3	50.8%
stative	15.4	**82.1**	2.4	42.8%
cannot decide	18.0	16.3	**65.7**	6.4%

(a) all segments

	dynamic	***stative***	***cannot decide***	**Frequency**
dynamic	**86.0**	11.9	2.2	64.6%
stative	25.1	**71.5**	3.4	30.0%
cannot decide	21.7	15.7	**62.6**	6.4%

(b) excluding segments where main verb is "have" or "be"

Table 6.10: Coincidence matrix for **lexical aspectual class**: pairings of labels by three annotators, in percent, normalized per row. MASC and Wikipedia data. **frequency** = percentage of all label assignments to segments.

mostly used the same label, and rarely ***cannot decide***. If one annotator used ***cannot decide***, in about one third of the cases, the other annotator found either the ***dynamic*** or ***stative*** reading predominant. For verbs other than "have" or "be" (see Table 6.10 (b)), confusion between ***dynamic*** and ***stative*** is higher; many cases labeled ***stative*** by one annotator are labeled ***dynamic*** by the respective other annotator.

	non-generic	***generic***	***expletive***	***cannot decide***	**Frequency**
non-generic	**90.8**	6.7	0.4	2.1	72.7%
generic	23.5	**74.2**	0.2	2.1	20.6%
expletive	58.4	0.9	**29.2**	3.5	0.5%
cannot decide	25.2	6.9	0.3	**67.6**	6.2%

Table 6.11: Coincidence matrix for **genericity of main referent**: pairings of labels by three annotators, in percent, normalized per row. MASC and Wikipedia data. **frequency** = percentage of all label assignments to segments.

Table 6.11 shows that annotators often disagree between ***non-generic*** and ***generic***, which hurts the less frequent category ***generic*** more. The only problematic category is ***expletive***, which has a high confusion rate with ***non-generic***.

This is due to cases such as (4), which were often labeled ***expletive*** by some annotators, where in fact the main referent "it" is just a cataphoric reference to the situation described thereafter, which is a particular ***non-generic*** situation. We instruct annotators to test whether "it" can be substituted with the *that*-clause and label the segment accordingly, but some annotators nevertheless frequently assigned ***expletive*** in such cases.

(4) It is clear (State, ***non-generic***)
that John is really interested in music. (State, ***non-generic***)

For lexical aspectual class and genericity of the main referent, the percentage of situation segments where annotators found it difficult to decide and chose ***cannot decide*** is between 6.2% and 6.4%, respectively. The percentage of items that received the label ***cannot decide*** by both annotators lies between 62.6% and 67.6%, indicating that at least to some extent, annotators have difficulties with the same items, which are on the boundary between the categories defined for each feature.

We study agreement for the habituality feature (see Table 6.12) on the subset of the segments that receives a situation entity label other than one of the subtypes of Speech Mode in the gold standard, as this feature does not apply for Questions and Imperatives.

	episodic	***static***	***habitual***	***cannot decide***	**Frequency**
episodic	**82.8**	12.5	3.5	1.2	28.5%
static	6.0	**87.7**	4.7	1.6	59.2%
habitual	9.3	26.0	**63.1**	1.6	10.8%
cannot decide	22.4	60.1	6.1	**11.4**	1.6%

Table 6.12: Coincidence matrix for **habituality**: pairings of labels by three annotators, in percent, normalized per row. MASC and Wikipedia data. **frequency** = percentage of all label assignments to segments.

According to the coincidence matrix for habituality in Table 6.12, the value ***cannot decide*** was used rarely; the values ***episodic***, ***habitual*** and ***static*** capture almost all cases. Agreement is good for ***episodic*** and ***static***, while there is a significant number of segments which received both the label ***habitual*** and the label ***static***. Cases that were labeled both ***habitual*** and ***static*** include main verbs similar to "be called" or "work at", and are related to disagreements on the level of lexical aspectual class.

6.3.4 Comparison of genres and categories

Inter-annotator agreement is not uniform across the various genres of MASC and the categories of the Wikipedia data. We here report agreement by genre and Wikipedia category, computing expected agreement according to the distributions in the respective subsets of the data. For lexical aspectual class, we use all situation segments regardless of the main verb's lemma. To compute agreement on habituality per genre, we again omit situation segments labeled as Question, Imperative or Speech Mode in the gold standard. From Tables 6.13 and 6.14, we can see that we achieve substantial agreement on lexical aspectual class, habituality and situation entity type across all MASC genres and Wikipedia categories. In contrast, the decision of whether the main referent is ***generic*** or not is apparently easier for some genres (ficlets or jokes) and very difficult in other genres (journal, travel guides and news). The technical texts contain cases that exemplify a concept using a hypothetical "concrete" example, e.g., talking about a particular robot which does not exist but which stands for any robot of the described kind. This caused many disagreements.

Genre	Main ref.	Asp. class	Habituality	SE type	Annotators
blog	0.49	0.66	0.59	0.62	A, C, D
email	0.58	0.68	0.62	0.65	A, B, D
essays	0.41	0.59	0.60	0.54	A, B, E
ficlets	0.76	0.78	0.77	0.80	A, C, D
fiction	0.59	0.70	0.73	0.77	B, C, D
govt-docs	0.41	0.59	0.61	0.57	A, B, D
jokes	0.74	0.79	0.74	0.77	A, B, D
journal	0.28	0.61	0.60	0.52	B, C, D
letters	0.57	0.69	0.61	0.66	A, B, C
news	0.33	0.67	0.73	0.75	B, C, D
technical	0.35	0.66	0.70	0.55	A, B, D
travel	0.32	0.66	0.70	0.59	B, D, E

Table 6.13: Inter-annotator agreement (Fleiss' κ) on features and situation entity (SE) type for MASC data.

In the Wikipedia data, agreement on the feature values for the main referent's genericity are much higher than for MASC due to the more balanced distribution of ***generic*** and ***non-generic*** cases. Categories with relatively low agreement comprise the texts about organized crime and ethnic groups; both contain texts about large but "limited" groups (gangs or tribes), and genericity needs to be determined in relation to the clause in each case, which makes annotation of these genres a difficult task (see Section 6.8 for a discussion).

The numbers presented in this section highlight which of the semantic phenomena modeled by our annotation scheme are reliably annotated in the respective parts of the corpus. Lexical aspectual class and habituality can be studied on any part of the corpus; when studying the subjects' genericity, using MASC alone is not sufficient as due to the low number of ***generic*** items in this part of the corpus, annotations are not reliable in about half of its genres.

Genre	Main ref.	Asp. class	Habituality	SE type
animals	0.62	0.60	0.60	0.64
biographies	0.61	0.66	0.66	0.63
botany	0.65	0.64	0.62	0.66
crime	0.54	0.70	0.75	0.66
ethnic groups	0.52	0.68	0.68	0.57
games	0.65	0.68	0.65	0.64
medicine	0.62	0.64	0.63	0.69
music	0.71	0.67	0.67	0.68
politics	0.61	0.56	0.57	0.59
religion	0.56	0.63	0.61	0.62
science	0.61	0.61	0.63	0.64
sports	0.70	0.66	0.69	0.69

Table 6.14: Inter-annotator agreement (Fleiss' κ) on features and situation entity (SE) type for Wikipedia data. Labeled by annotators A, B and C.

6.3.5 Agreement on Abstract Entities

Abstract Entities are the situation entity types including Facts, which represent assessments of states of affairs, and Propositions, which are objects of mental states such as belief, expectations and decisions (Smith, 2003, p. 74). For details, see Section 5.2.3. We ask our annotators to label situation entities as Fact or Proposition where appropriate, and give the basic underlying situation entity type of the Abstract Entity in addition, as illustrated in example 5.

(5) I think (State)
that he will take this job. (Proposition, basic situation entity type: State)

In general, our data contains only very few Abstract Entities. We analyse agreement on detecting Abstract Entities in the same way as we analyze agreement on correcting the automatic segmentation above, giving the percentage of segments that are labeled as an Abstract Entity by one, two or three annotators (see Table 6.15).

		# annotators		
Corpus	# situation segments	3	2	1
MASC	38025	0.4	0.7	2.1
Wikipedia	11984	0.0	0.5	1.5

Table 6.15: Observed agreement in percent on whether a segment is labeled as Abstract Entity by 3, 2 or one annotator.

On the Wikipedia data, a large subpart of our corpus which has been labeled by the same three annotators, the number of Abstract Entities detected by each annotator ranges from 11 to 174. The annotator that used Abstract Entity most includes complements of verbs like *conclude*, *there is evidence*, *it is possible*, *say* or *suspect*. Some of these cases are on the borderline with the definition of Abstract Entities as expressing states of the mind, some, like *it is possible*, should probably not be considered to introduce Abstract Entities.

While reaching a good recall of Abstract Entities on our corpus proved difficult, agreement on the subtypes of Abstract Entity is good. Table 6.16 compares the cases that received an Abstract Entity by two or more annotators, we compare how often they agreed on whether the relevant segment is a Fact or a Proposition. The confusion matrix presented in Table 6.16 is based on *unordered* pairs of label assignments (hence there is no entry of Fact-Proposition).

	Fact	Proposition
Proposition	18	238
Fact	54	-

Table 6.16: Confusion matrix for subtypes of Abstract Entity on MASC and Wikipedia.

The 18 disagreements between Fact and Proposition contain some apparent attention slips, but also cases such as (6) where it is indeed not trivial to interpret whether the author presents the embedded situation as a fact or a belief, or where additionally the thought is attributed to another party (7).

(6) It seems clear (State)
that bracket matching can't in general be required. (Abstract Entity)

(7) He was sure (State)
his Mother was a virgin. (Abstract Entity)

6.4 Annotator bias

One factor that affects reliability is the *individual annotator bias*, i.e., the individual preferences of an annotator for choosing the various categories of an annotation scheme. The development of clear annotation schemes and guidelines aims at avoiding annotator bias; however, some individual differences in the interpretation of the manuals will always remain (Artstein and Poesio, 2005). Modeling annotator bias has been shown to be helpful when compiling a gold standard from crowd-sourced annotations (Klebanov and Beigman, 2009; Passonneau and Carpenter, 2014). As we have worked with three expert annotators per corpus section, we construct our gold standard simply via majority voting. Nevertheless, we evaluate annotator bias in this section with the aim of understanding whether an annotator uses a label more (or less) often than his or her colleagues in general, i.e., whether he or she has a individual bias concerning this label. This, in turn, helps to estimate which parts of our annotation manual and guidelines were clear, and which parts may need refinement in future work.

Our annotators worked on the different sections of our corpus in various combinations. The different corpus sections may have different "true" distributions of the various labels, so we cannot simply look at the statistic how often an annotator used a label in comparison to his or her colleagues. Instead, we compute a statistic that can be interpreted as the percentage to which an annotator over- or under-uses a label in relation to how often it is used on average by the other annotators.

Following the notation introduced in Section 6.3.2, we use $n_{k,s}$ to indicate the number of times label (or category) k has been assigned to an item in section s. Corpus section s has i_s items which all have been labeled by c_s annotators (coders), so the average frequency with which label k has been used in s is:

$$\mu_{k,s} = \frac{n_{k,s}}{i_s c_s}$$

and the frequency with which annotator a has used k in s is:

$$\mu_{k,s,a} = \frac{n_{k,s,a}}{i_s}$$

We are interested in whether annotator a used label k more or less frequently than it has been used for corpus section s on average. This difference is expressed by $\mu_{k,s,a} - \mu_{k,s}$, which is positive if an annotator used a label more frequently than average, and negative otherwise. If labels are applied only infrequently, even small deviations may be meaningful. For this reason, we divide the resulting difference by the mean, resulting in a bias score per annotator, label and corpus section:

$$bias_{k,s,a} = \frac{\mu_{k,s,a} - \mu_{k,s}}{\mu_{k,s}}$$

Finally, we average this number over all corpus sections in which an annotator participated, obtaining a bias score $bias_{k,a}$ per annotator and label. We treat the MASC genres and the entire Wikipedia data as corpus sections. If noise was random, this $bias_{k,a}$ should be close to zero for all annotators and labels; if annotators have biases, we should see positive or negative numbers.

$$bias_{k,a} = \frac{\sum\limits_{s \in \{a\ marked\ s\}} bias_{k,s,a}}{|s \in \{a\ marked\ s\}|}$$

Our analysis as presented in Table 6.17 offers a high-level view of annotator bias. Note that values with a large absolute value can be caused by a a difference in understanding compared to another annotator who marked the same section(s). The differences shown in Table 6.17 depend on each other. For example, if one annotator heavily under-uses a category, this will result in positive values for some of the other annotators. The numbers do not allow for judging which annotator is "right", and they are less reliable for annotators who only marked few corpus sections, but they offer a starting point for analyzing individual annotators' behaviors.

Task	Label	A	B	C	D	E
situation entity type	State	-0.6	6.43	-8.22	-1.59	4.07
	Event	-1.84	4.56	13.79	-14.78	8.91
	Report	-15.82	39.88	16.5	-3.4	-4.56
	Generic Sentence	22.46	-9.61	-33.94	25.12	-55.01
	Generalizing Sentence	-21.2	12.15	10.28	-10.98	47.51
	Question	-13.39	25.94	7.07	-0.92	-6.04
	Imperative	5.33	-4.04	7.98	-1.36	-22.89
	None	2	-18.89	13.1	8.18	8.18
lexical aspectual class	***dynamic***	-7.54	11.89	10.02	-15.48	10.85
	stative	7.23	-6.33	-14.27	13.46	-15.12
	cannot decide	0.94	-16.41	6.82	9.42	15.01
main referent	***non-generic***	-3.17	6.48	1.32	-7.03	9.16
	generic	20.77	-6.1	-34.29	22.75	-53.65
	expletive	-6.02	-74.84	53.55	50.41	34.3
	cannot decide	1.16	-16.03	8.11	9.06	9.27
habituality	***episodic***	-4.74	6.31	13.04	-13.36	7.79
	habitual	-5.13	3.58	4.75	-6.07	17.13
	static	2.25	2.55	-10.99	4.57	-8.53
	cannot decide	1.96	-14.75	10.01	5.44	10.03
number of categories annotated		9	11	7	10	2

Table 6.17: Annotator bias $bias_{k,a}$. Darker cells indicate higher percentage-wise deviations from mean.

On the level of situation entity types, annotator E assigned Generic Sentence less frequently than the annotators marking the same data sets. This is matched by the observation that the same annotator used ***generic*** less frequently. Annotator C has the same tendencies as annotator E, but less pronounced. Annotator D has the opposite tendency. We also observe clear tendencies for lexical aspectual class: annotators B, C and E tend towards using ***dynamic*** more often, while annotators A and D have a preference for using ***stative*** more often. From this chart, we can also tell that B, C and D marked ***expletive*** more often than average, while B used it less often. Annotator B also uses ***cannot decide*** less often for all annotation tasks, thus giving labels in cases where her co-annotators refuse to give one.

To summarize, as already pointed out in Section 6.3.2 and Section 6.3.3, the major disagreements are related to genericity (Generic Sentence and ***generic***). In this section, we have shown that these disagreements involve individual preferences, i.e., annotators in fact apply slightly different understandings of these categories. In this section, we have taken a detailed look at how annotators differ from *each other*; in the next section, we investigate how consistently they behave with regard to their *own* annotations over time.

6.5 Intra-annotator agreement

The aim of the experiments presented in this section is to establish the degree of noise contained in the data. We measure *intra-annotator agreement*, i.e., the extent to which annotators reproduce their own decisions.

For this study, we chose approximately 200 situation segments per MASC genre and 515 situation segments from the Wikipedia part of the data. We use texts (or beginnings of texts) from our corpus that have a low or medium level of agreement in order to measure the intra-annotator stability on the difficult rather than the easy decisions. We ask annotators to re-annotate some of the documents several (> 3) months after they annotated them for the first time, without having access to the annotations created in the first round.

We compute intra-annotator agreement as κ for each annotator, comparing the annotations an annotator gave in the first and second round of annotation. As in Section 6.4, the various annotators have marked different sections. We here need to take into account that some sections seem to be more difficult to annotate than others (see Section 6.3.4). For each section and round, we compute inter-annotator agreement between the three annotators who marked the section. We average these numbers, as well as the intra-annotator agreement κ scores, for each annotator over the sections that he or she has marked; these numbers are shown in Table 6.18. Interestingly, we observe that inter-annotator agreement is higher in the second round than in the first round in almost all cases. A possible explanation is that the longer annotators are involved in the project, the more familiar they become with the guidelines and thus the more consistent are the annotations that they produce.

Intra-annotator agreement is, in each case, much higher than the corresponding inter-annotator agreement: the noise that we observe *between* annotators is more than that we observe *within* annotators. In other words, annotators reproduce their own decisions more often than they reproduce other annotators' decisions. Most of the intra-annotator agreement scores for situation entity types, lexical aspectual class and habituality are in the range of substantial and almost perfect agreement (Landis and Koch, 1977), ranging from 0.67 to 0.87 (with the exception of annotator E, discussed below). Labeling the genericity of the main referent is again the most difficult task; the scores of three annotators are in the substantial range here, while the other two annotators reach only moderate agreement with themselves.

Annotator E was only involved in two genres, which turned out to be two of the most difficult ones to annotate. In the travel genre, it was often difficult to decide whether a reference to a nation (e.g., "*the British*") is generic or not; the essays genre is difficult due to the often abstract concepts mentioned in the texts. The intra-annotator agreement scores for the individual annotators are not directly comparable as they were involved in different sets of sections, and for some genres it is easier to obtain high κ scores than for others. In order to estimate to what extent annotators reproduce their own decision while abstracting away from the difficulty of the genres that were assigned to them, we compute a score $intra_{diff}$ (see Table 6.18). For each row shown in Table 6.18, we compute the inter-annotator agreement κ_{inter} as the average of the scores reached in the first

Task	**Annotator**	**# genres**	**first**	**second**	**intra**	$intra_{diff}$
SE type	A	9	0.59	0.67	0.82	31.5
	B	11	0.57	0.61	0.67	12.9
	C	7	0.62	0.67	0.81	25.1
	D	10	0.61	0.66	0.71	12.1
	E	2	0.49	0.46	0.56	17.7
lexical	A	9	0.65	0.64	0.87	35.0
aspectual	B	11	0.63	0.63	0.76	20.3
class	C	7	0.68	0.69	0.85	24.4
	D	10	0.65	0.65	0.81	23.5
	E	2	0.65	0.66	0.82	24.6
main referent	A	9	0.55	0.59	0.79	38.2
	B	11	0.50	0.55	0.58	9.6
	C	7	0.60	0.56	0.75	29.5
	D	10	0.52	0.55	0.68	26.8
	E	2	0.40	0.38	0.45	17.0
habituality	A	9	0.60	0.66	0.82	30.6
	B	11	0.59	0.63	0.69	14.3
	C	7	0.63	0.64	0.83	30.3
	D	10	0.62	0.66	0.76	18.9
	E	2	0.56	0.56	0.71	27.2

Table 6.18: Intra-annotator agreement. The columns labeled **first** and **second** show inter-annotator agreement of the first and second annotation rounds respectively in terms of Fleiss' κ. The column labeled **intra** shows the intra-annotator agreement per annotator (also using Fleiss' κ).

and second annotation rounds. The score $intra_{diff}$ indicates (in %) to what extent an annotator reproduced their own decisions compared to the how well annotators produced each other's decisions (reflected by κ_{inter}) in the respective relevant set of corpus sections:

$$intra_{diff} = \frac{intra - \kappa_{inter}}{\kappa_{inter}}$$

The logic behind the weighting scheme is similar to the procedure as described in Section 6.4 or as in computing Cohen's κ from observed and expected agreement.

We observe that this percentage-wise estimate of intra-annotator agreement is not directly correlated to the absolute κ values for each annotator, annotator E reaching intermediate values when compared to the entire set of annotators. Annotator A is the one that has the highest stability in reproducing own decisions followed by annotator C; annotator B is the one that seems least stable.

We have learned from this study that annotators reproduce their own decisions with satisfying quality in most cases. As intra-annotator agreement scores are not perfect in several cases, the data also contains some noise due to randomness of annotator decisions. The difficult cases here are the same as the ones observed for inter-annotator disagreement. For lexical aspectual class and habituality, intra-annotator agreement is substantial to almost perfect; for situation entity types and main referents, some texts with instances that are unclear with regard to our genericity annotation guidelines cause lower scores.

6.6 Gold standard construction

We construct the gold standard for our experimental evaluations (see Part III) using majority voting over the annotations given by three annotators. This section describes how we create the gold standard from the set of annotations given for each segment.

6.6.1 Segmentation

The first step in creating a gold standard for situation entity types from the provided annotations is to decide whether to consider each of the automatically created segments as invoking a situation entity. Annotators indicate whether they consider something to be a situation entity or not, but they do not necessarily agree on this decision (see section 6.3.1). For the construction of our gold standard, we use all segments that receive a situation entity type label by at least two of the three annotators.[5]

6.6.2 Situation entity types

We create our gold standard of situation segments labeled with their situation entity types using majority voting over the following labels:

- State
- Event
- Report
- Generic Sentence
- Generalizing Sentence
- Imperative
- Question

This list does not include Abstract Entities because they have not been annotated reliably in our corpus (see Section 6.3.5). Annotations for Abstract Entity and its subtypes lack in recall, so simply using majority voting over the three annotators would not result in a good gold standard. However, as illustrated by (8), situation segments labeled as Abstract Entity in addition receive a label of one the situation entity types or subtypes of State, Event or General Stative. We use these labels for our gold standard, treating

[5] The remaining segments are not labeled as a situation entity and excluded from our experiments. However, we distribute these unlabeled segments along with our corpus for potential future use.

Abstract Entities completely separately from the classification of segments into the above set of situation entity types (see Section 8.5.6 for experiments).

(8) (a) I know (State)
(b) that you want it. (Fact / State)

For Questions and Imperatives, we do not ask the annotators to mark one of the other SE types. For the gold standard construction, if two or more annotators agree on one of the Speech Mode types, we use this type as the segment's gold standard label.

The SE type Report is a subtype of Event intended to mark attribution or events denoting acts of speech introducing direct or indirect speech. As discussed in Section 6.3.1, in many cases, our annotators neglect to mark Report and mark the segment as Event instead. If at least one annotator identified a Report, and the other two annotators marked the situation entity as Event, most cases actually should be labeled Report according to the author's judgment. For this reason, we decide to use the gold standard label Report if at least one annotator marked a situation as such, and the other annotators marked it as Event. The majority of these cases are like example (9): there are no segmentation errors, and segment (a) has been labeled as Event by two annotators and as Report by the third.

(9) (a) He told them (Report)
(b) he didn't know. (State)

There are a few cases involving segmentation errors such as example (10), which contains two situation entities within one automatically created segment.

(10) "I'm just not sure that I could," he said. (Report)

Annotators are instructed to provide labels for the clause that would end up higher in a dependency parse in such cases (as illustrated in Figure 6.1), i.e., for the clause *he said* in this case. Unfortunately, annotators did not always annotate these cases consistently and occasionally provide labels for the situation entity enclosed by the quotation marks. This is most likely a result of our instruction to annotate the first situation entity in case of multiple situations in the general case (see Section 6.3.1). Inspired by several examples such as (10), we decide to apply our rule-of-thumb to use Report as the gold standard label for situation entities that receive this label at least once also in this case.

One percent of the segments received labels that allowed the inference that the segment is a General Stative, but no agreement was reached on whether it is a Generic Sentence or a Generalizing Sentence. These are cases where annotators agree that the clause is ***habitual***, but disagree on whether the main referent is ***generic*** or not. These cases are labeled as General Stative in the gold standard, but we did not treat them as their own category in the inter-annotator agreement analysis because this label is a super-category.

As explained in Section 6.3.1, annotators can refrain from giving a situation entity type label for a segment, and we include all segments that received a situation entity type label by at least two annotators into the gold standard. There are only very few (2.4%) segments

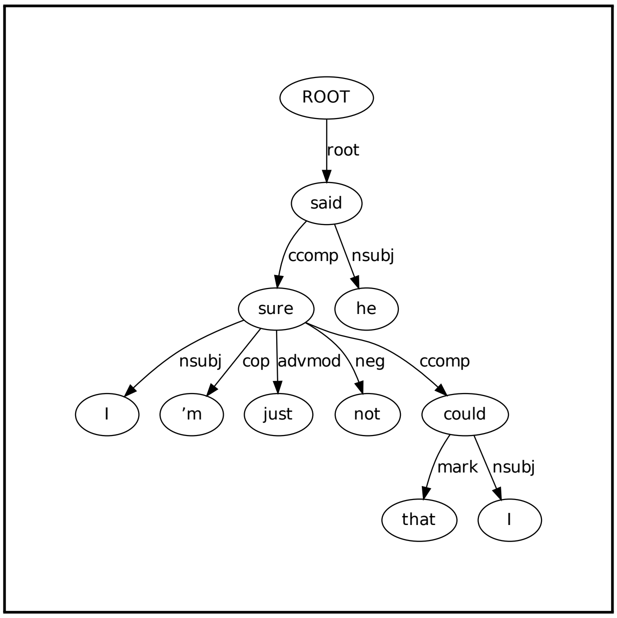

Figure 6.1: Dependency parse tree for example (10), using Stanford typed dependencies (de Marneffe and Manning, 2008a).

that receive two situation entity type labels, but where, according to rules consistent with those explained above, majority voting does not result in one of the above types. In 48 cases containing speech acts of thanking, one annotator used the label Speech Mode, while one other annotator used Event or State. For cases where no majority vote could be reached, the set of situation entity types assigned to the segment contains Speech Mode and the segment's text contains "thank", we set the situation entity type to Speech Mode. We do not use Speech Mode as a separate label in the agreement analysis nor in the automatic classification experiments due to its use as a super-category of Question and Imperative and due to the low frequency of direct assignments of this label.

Finally, there are segments that receive two or more incompatible situation entity type labels by at least two annotators. Simply omitting these instances from the gold standard does not seem right, as they are different from the segments not invoking a situation entity at all. They *do* invoke a situation entity, but they constitute the very hard cases on which not even a majority of the annotators could agree. We use the label CannotDecide for these cases. Lexical aspectual class can often be interpreted in different ways in these cases as in (11), which has been labeled as Event and State, while the third annotator refused to give a label. Other cases include disagreements on the level of habituality or genericity in addition; (12) has been labeled as Event, State and Generalizing Sentence by the annotators and (13) as State, Generic Sentence and Imperative.

(11) [An angel,...] fallen to the earth (CannotDecide) (MASC blog)

(12) Detroit police are watching you! (CannotDecide) (MASC ficlets)

(13) You must never confuse faith (CannotDecide) (MASC email)

6.6.3 Situation-related features

Our gold-standard for the situation-related features is created using majority vote over the set of labels presented in Table 6.8. We apply some post-processing steps, which make up for some of the problems that occurred during annotation as reported above. These post-processing steps affect rare cases but lead to a better gold standard.

If no agreement could be reached on the genericity of the main referent, this could be due to the combination of labels ***generic*** – ***non-generic*** – ***expletive***. As explained in Section 6.3.3, some annotators had difficulties distinguishing actual expletive uses from cataphoric ones. Hence, if majority voting does not result in assigning something other than ***cannot decide***, and the set of labels contains ***expletive***, we correct this annotator's vote to ***generic*** or ***non-generic*** according to the situation entity type label assigned by this annotator – if this label is Generic Sentence, we assign ***generic***, and ***non-generic*** otherwise. Then, majority voting is attempted again.

During some time of our annotation project, we allowed the label ***both readings*** for lexical aspectual class. However, we also have the label ***cannot decide***. Cases labeled as ***both readings*** are in fact mostly cases that should have received the label ***stative***, as in example (14).

(14) Linguistic categories such as situation entity types are called "covert."

If no label could be assigned via majority voting, we attempt majority voting again, changing votes for ***both readings*** into ***stative***.

Finally, despite the instruction not to give values for the habituality feature in the case of Speech Mode types, sometimes annotators do so. We do not assign values to the habituality features to segments labeled as Imperative or Question in the gold standard.

6.7 Label distributions in gold standard data

We conclude the presentation of our corpus of MASC and Wikipedia texts annotated for situation entities, and related features by showing the distribution of labels as marked in the gold standard. The differences between these distributions highlight variations between genres (see also Palmer and Friedrich, 2014), and in addition, awareness of these differences is necessary for interpreting the experimental results in Part III.

Situation entity types. Figures 6.2 and 6.3 show the overall distribution of situation entity types in MASC and the Wikipedia data in terms of absolute situation entity counts. Figure 6.4 gives the statistics per genre for MASC, and figure 6.5 for the Wikipedia corpus.

The cases listed as General Stative are the cases where the majority of annotators agreed on the supertype General Stative, but no majority agreement was reached on whether the situation entity should receive the label Generic Sentence or the label Generalizing Sentence. Cases labeled as one of the latter two types according to the gold standard are General Statives, too, but their instances are not added to the statistics of General Stative.

As explained in Section 6.6, there are 48 cases labeled as Speech Mode in the corpus. They all occur in the MASC part of the data; mostly in email, (fund-raising) letters and jokes.

The predominant situation entity type in the MASC data is State, followed by Event. Other types are much less frequent, though the technical, essays, jokes and blog genre all contain between 11.3% and 16.8% Generic Sentence. Genres containing narratives (ficlets and fiction) have the largest percentage of Events, which is in line with the intuition that States and Events are the predominant situation entity types in the **Narrative** mode.

As stated in section 6.1.2, this set of Wikipedia documents was collected in order to study the phenomenon of genericity on a dataset with a sufficient, but not overwhelming number of generic sentences. The categories differ with respect to the percentage of Generalizing Sentences (see table 6.5): botany, games, animals, music and medicine have more generic than non-generic sentences; crime, ethnic groups, politics, religion, science and sports have distributions with between 25 and 50% generics, and biographies are a more ‘narrative’ genre with only 7.4% Generic Sentences and many Events. Interestingly, the biographies category also has the highest percentage of Generalizing Sentences (4.9%).

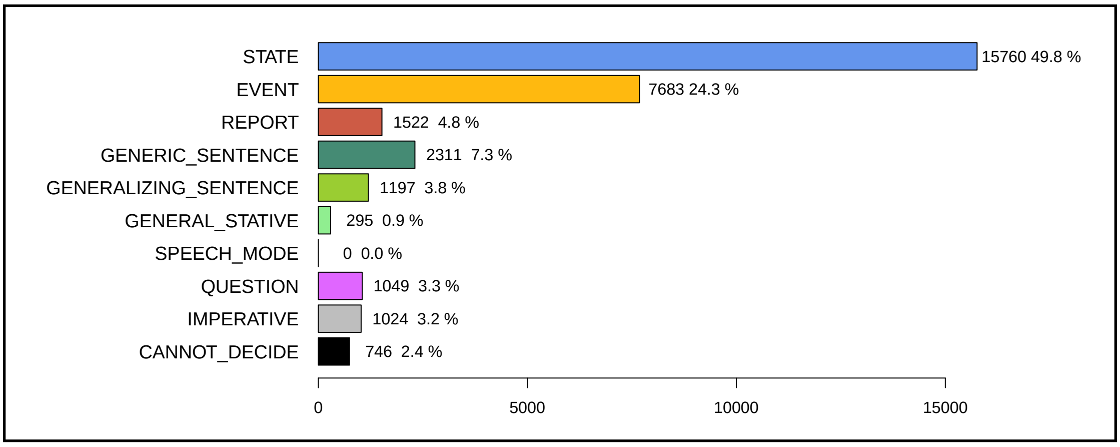

Figure 6.2: Distribution of situation entity types: MASC, absolute counts and percentages.

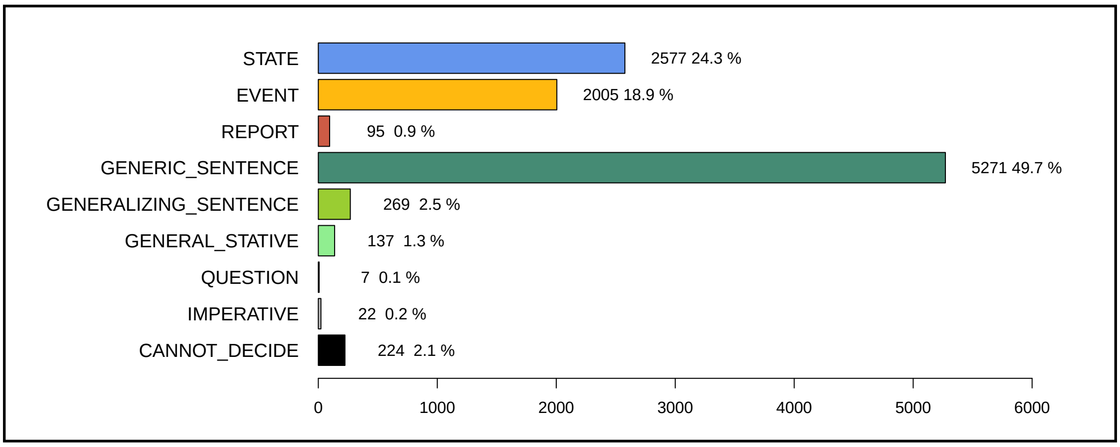

Figure 6.3: Distribution of situation entity types: Wikipedia, absolute counts and percentages.

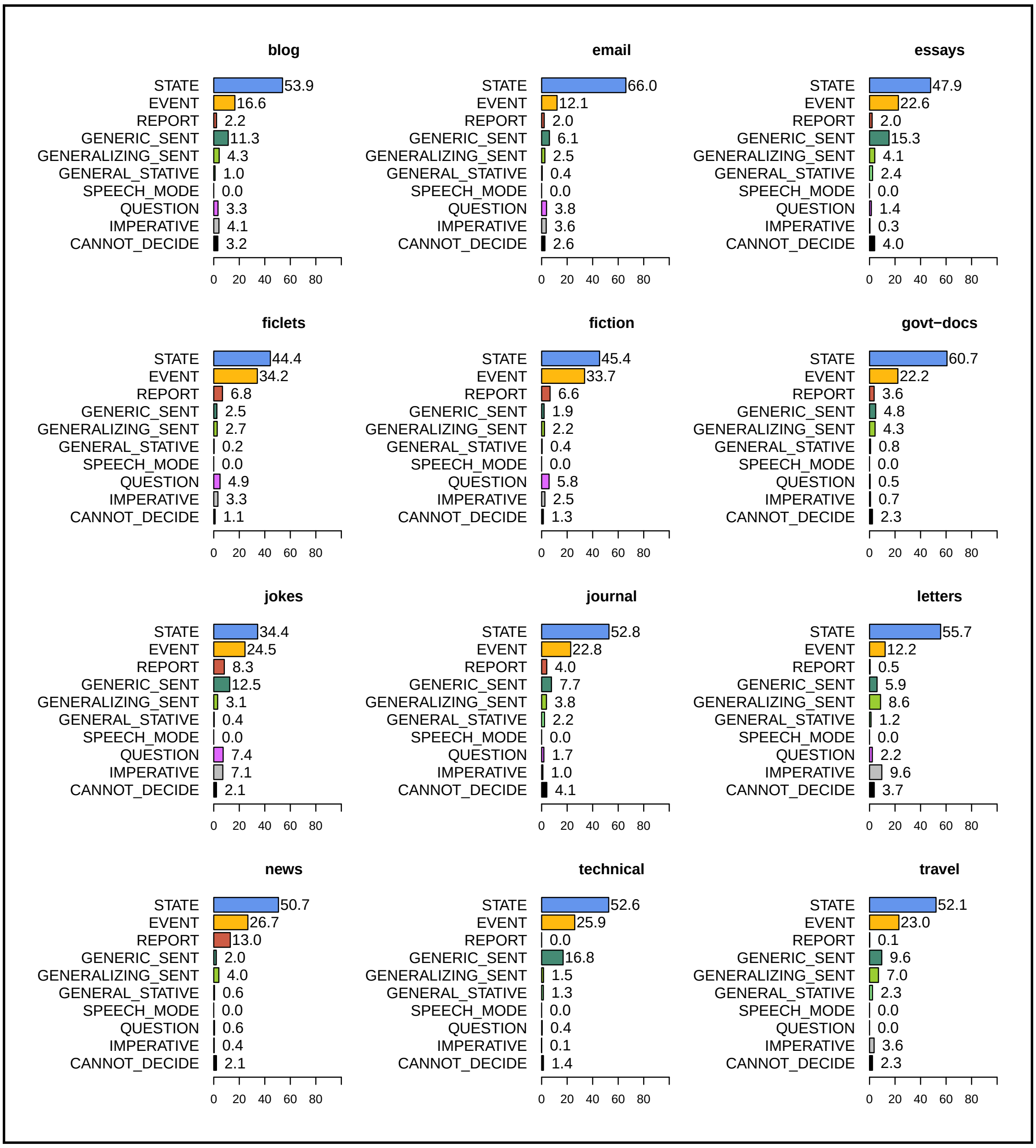

Figure 6.4: Distribution of situation entity types in MASC: normalized per genre.

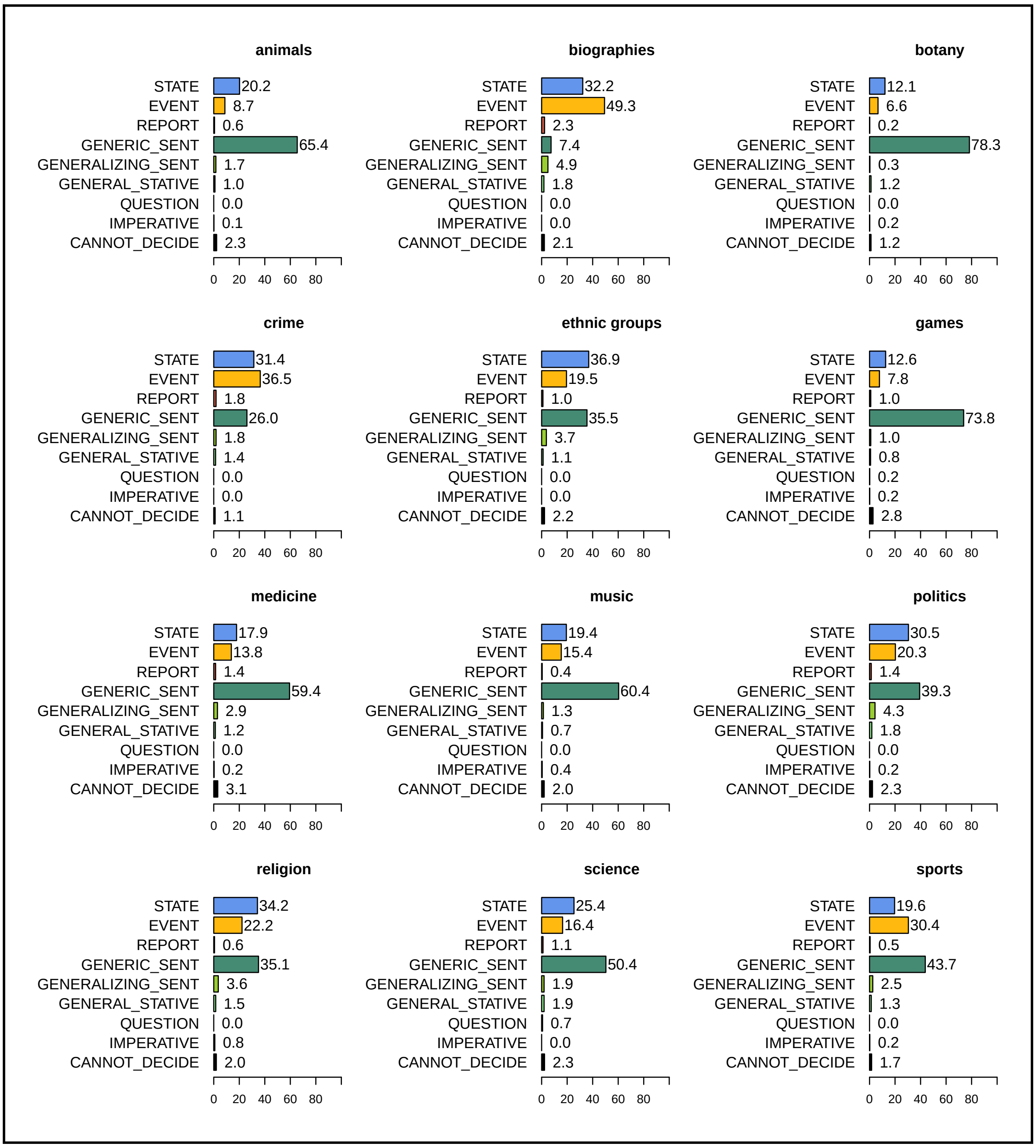

Figure 6.5: Distribution of situation entity types in Wikipedia: normalized per category.

Situation-related features. Tables 6.19, 6.20 and 6.21 show the distributions of the situation-related features' values in the gold standard. The percentage of ***dynamic*** cases is approximately the same in MASC and Wikipedia. The three annotators marking Wikipedia almost never used ***cannot decide***, while this label was used in 7.5% of the MASC data. This is in line with our assumption that most of the cases labeled as **both readings** are actually ***stative*** cases.

Label	MASC	Wikipedia
dynamic	51.2	50.3
stative	41.3	48.9
cannot decide	7.5	0.8

Table 6.19: Lexical aspectual class: distribution of labels in gold standard.

The cases listed as *n/a* for habituality are the percentage of segments that are labeled as Question, Imperative or Speech Mode in the gold standard. Wikipedia contains more "informative" text, and hence more ***generic*** and ***habitual*** cases than the MASC part of the corpus.

Label	MASC	Wikipedia
episodic	29.3	20.1
habitual	6.1	19.5
static	55.8	58.6
cannot decide	2.0	1.5
n/a	6.7	0.3

Table 6.20: Habituality: distribution of labels in gold standard.

Label	MASC	Wikipedia
non-generic	84.1	43.1
generic	7.9	55.3
expletive	0.4	0.0007
cannot decide	7.6	1.6

Table 6.21: Genericity of main referent: distribution of labels in gold standard.

6.8 Discussion

The annotation scheme as presented in chapter Chapter 5 is guided by the framework of situation entity types as suggested by Smith (2003). In order to understand the major differences between the situation entity types she defines, we employ the three distinctions of lexical aspectual class, genericity of the main referent and habituality of the situation entity. This is helpful in the majority of cases and allows annotators to gain an understanding of the scheme and guidelines quickly, and to make consistent decisions. However, during the agreement study and gold standard construction presented in this chapter, we identified several cases that are not easily classified by the three linguistic features according to which we classify situation entity types. In this section, we present a critical analysis of these cases, highlight the limitations of the current annotation guidelines, and identify potential improvements for future work.

6.8.1 Lexical aspectual class and habituality

The aspectual type of a proposition may change under the influence of modifiers such as tense, temporal adverbials or aspectual adverbials (Moens and Steedman, 1988). Example (15) illustrates the various interpretations a verb can take in different contexts.

(15) (a) He is **eating** a sandwich. (***dynamic***, ***episodic***)
(b) He usually **eats** meat on Sundays. (***dynamic***, ***habitual***)
(c) Is he vegetarian? - No, he **eats** meat. (***stative***, ***static***)

(15c) is interpreted as ***stative*** as it ascribes a property to the subject, namely that this person is a meat-eater. However, this word sense of "eat" still reflects its etymologic source, which includes a habitual use of "eat" (see also Section 2.1.1). There is no clear boundary between such cases, where the lexical entry can be considered as stative, and cases that are simply habituals. In the first phase of our annotation project, we gave only few examples of these difficult cases that are to be considered as ***stative*** ("work at", "study at"). Frequently occurring cases on which annotators disagreed include "call" and "refer to". In a later iteration of our annotation manual, we instruct annotators to treat these cases, which primarily ascribe properties, as ***stative***.

In addition, annotators sometimes misinterpret clauses including the word *often*, generalizing only over the subject noun phrase, as generalizing over situations. The clause in (16b) actually says that *many* of the earlier factions were tied to particular leaders, not that this happened repeatedly.

(16) (a) As opposed to the instability of the earlier factions, (State)
(b) which were often **tied** to a particular leader (State)
(c) [...] the party was centred around a set of core principles (State).

(16b) is one of the cases receiving the label General Stative in our gold standard because it was labeled as a Generic Sentence, a Generalizing Sentence and a State by one annotator each. The author of this thesis analyzes these examples as a State.

6.8.2 Genericity

In many cases, distinguishing between kind-referring (***generic***) and non-kind-referring (***non-generic***) noun phrases is a challenging task for annotators. Example (17) shows three sentences taken from the Wikipedia entry on the Ayoreo people. In sentence (a), the first sentence of this text, it is already underspecified whether *The Ayoreo* refers to a kind or to a small particular tribe. Sentences (b) and (c) differ with regard to genericity: (b) says that any person who is an Ayoreo has a certain lifestyle, therefore the subject noun phrase is generic, and (c) says that the particular tribe of Ayoreo (which the reader has learned to be small at this point) is threatened.

(17) (a) The Ayoreo are an indigenous people of the Gran Chaco. (Generic Sentence / State)
(b) Ayoreo combine hunter-gatherer lifestyle with farming. (Generic Sentence)
(c) The Ayoreo are threatened by deforestation. (State)

Besides such cases that are simply hard to annotate, but for which we offer an analysis according to the annotation guidelines, there are cases that are not well captured by our annotation guidelines. Example (18), the definition of the term "confederation", is such a case.

(18) "Confederation" refers to the process of (or the event of) establishing or joining the Canadian federal state. (State)

Intuitively, this sentence expresses generic knowledge and should therefore receive the label Generic Sentence. However, the particular term "confederation" is not kind-referring (disregarding the linguistic distinction of types and tokens), so annotators mark the main referent of this situation entity as non-generic, which in turn leads to the situation entity being labeled as a State rather than a Generic Sentence. This raises the question of whether the definition of Generic Sentence in our guidelines should be augmented by including any definitions expressing world knowledge, even if the main referent does not strictly refer to a kind.

Example (19) is an excerpt from the Wikipedia entry on trees. Disagreement occurred in the case of (19c), which was labeled as State, Generic Sentence and Generalizing Sentence. The clause receives the label General Stative in the gold standard.

(19) (a) They have actinorhizal root nodules on their roots (Generic Sentence)
(b) in which the bacteria live. (Generic Sentence)
(c) This process enables the tree to live in low nitrogen habitats (General Stative)
(d) where they would otherwise be unable to thrive (Generic Sentence).

Two annotators regarded *this process* as ***non-generic***, one annotator labeled it as ***generic*** (probably assuming that each actual tree has its own instance of this process). Two annotators considered *enables* as stative, and one annotator (the one who gave the label Generalizing Sentence) considered it as ***dynamic*** and ***habitual***. All of these interpretations are at least to some extent comprehensible.

Smith (2003, p. 72) defines General Statives as "being more abstract than States, because they do not express particular situations." She then defines General Statives as Generic Sentences and Generalizing Sentences, giving examples that are in line with our definitions in section 5.1.3. After conducting the annotation and evaluating agreement, we now propose to clarify to annotators that there are also some Generic Sentences whose subject noun phrase is not necessarily kind-referring, such as *this process* in (19). Arguably, these cases are *not* States in the sense that they express 'particular situations'.

Future work. In the current work, we have focused on NP-level genericity in order to distinguish Generic Sentence from other situation entity types. However, genericity involves other semantic phenomena as well. Krifka et al. (1995) distinguishes *habitual* and *lexical characterizing* sentences. Habitual sentences are labeled as either Generalizing Sentence or Generic Sentence in our annotation scheme, and lexical characterizing sentences with a ***generic*** subject are labeled as Generic Sentence. Lexical characterizing sentences with ***non-generic*** subjects (20), in contrast, are labeled as State in our scheme, just as any other state (21).[6]

(20) Simba has a mane.

(21) Simba is in this cage.

From the perspective of discourse modes, it might be worthwhile to capture the difference between (21), which gives information on a particular situation and which would probably occur in **Narrative** or **Report** mode, and (20) which gives background information and is more likely to occur in **Description** or **Information** mode.

Finally, example (22) (see also Section 5.2.2) constitutes a case where deviating from our strict definition of interpreting the subject's genericity in order to determine a situation entity's type might be necessary.

(22) <u>Potatoes</u> are served whole or mashed as a cooked vegetable.

(22) is labeled as Generic Sentence according to our guidelines, but it neither makes a statement about all potatoes nor about a typical potato. Rather, the context being the description of a restaurant, it says how potatoes are served in this place. The semantics of these sentences thus seem to be better captured by Generalizing Sentence. In future work, we plan to investigate to what extent the definition that a clause's subject is automatically the situation's main referent can be softened.

6.9 Summary

We have created a corpus of texts taken from MASC and Wikipedia annotated for their situation entity types, corresponding to approximately 40,000 labeled situation segments. The segments have additionally been labeled for the lexical aspectual class of their main verb, habituality and the main referent's genericity.

Inter-annotator agreement on situation entity types and the situation-related features is substantial, with genericity being the most difficult distinction to make for annotators.

[6] Examples by (Krifka et al., 1995, p.18).

Annotators exhibit different biases in the frequencies with which they make use of particular labels. The disagreement concerning genericity is thus not random noise, but results from slightly different understandings of the guidelines. The annotation of Abstract Entities lacks in recall. Our analysis of the various genres suggests that our annotation guidelines are easier to apply on some genres than on others.

We have also reported on details concerning the gold standard construction via majority voting and given an overview of the label distributions in the gold standard. The analyses conducted in this chapter form the basis for a future refinement of the annotation scheme and guidelines and for interpreting the results of our computational models as presented in the next part of this thesis.

Part III

Methods and experimental evaluation

Chapter 7

Computational modeling

This chapter explains our computational models for situation-related features and situation entity types. We approach all of these tasks in a supervised classification setting, using either classifiers labeling single instances or a sequence labeling method. We first explain the set of features that we use to describe the instances to be classified (Section 7.1). These features represent the data when training and testing the classification methods (Section 7.2). The various combinations of features and methods that we use in our experiments are described in Chapter 8. The implementation of the features and the code for training and testing models is freely available.[1]

7.1 Features for data representation

We represent each instance, i.e., each situation segment, its main verb, and the NP denoting its main referent, using a range of syntactic-semantic features. Table 7.1 gives an overview of all features.

7.1.1 Preprocessing

Texts are pre-processed with Stanford CoreNLP (Manning et al., 2014), including tokenization, POS tagging (Toutanova and Manning, 2000; Toutanova et al., 2003), lemmatization and dependency parsing (Klein and Manning, 2002) using the UIMA-based DKPro framework (Ferrucci and Lally, 2004; Eckart de Castilho and Gurevych, 2014).

7.1.2 Extraction of main verb and main referent

As explained in Section 5.1.1, the main verb and its subject (the main referent) carry important information with regard to a situation entity's type. Situation segments are given in our corpus as spans of text; segmentation is performed automatically as detailed in Chapter 4. We first determine the non-auxiliary verb within this span that is ranked

[1] `http://www.coli.uni-saarland.de/projects/sitent`

set	explanation	features/examples
pos	POS tags	binary: whether POS tag occurs in the segment
bc	Brown clusters	110111:2 if cluster 110111 occurs 2 times in the segment
mv	features describing the **main verb** & its arguments	tense, lemma, lemma of object, auxiliary, WordNet sense and hypernym sense, progressive, POS, perfect, particle, voice, linguistic indicators
mr	features describing the **main referent**, i.e., the NP denoting the main verb's subject	lemma, determiner type, noun type, number, WordNet sense and supersense, dependency relations linked to this token, person, countability, bare plural
cl	features describing entire **clause** that invokes the situation entity	presence of adverbs / prepositional phrases, conditional, modal, whether subject before verb, negated, verbs embedding the clause

Table 7.1: Overview of feature sets.

highest in the dependency parse of the sentence covering the span; this is the situation entity's main verb. We then specificy the grammatical subject of the main verb as the main referent. While the main verb must occur within the clause, the main referent may be a token either within or outside the clause's span. The sentence in example (1) has two clauses. "John" is the subject of each clause, despite not occurring in the span of the second clause. In the latter case, it still functions as the clause's main referent, as in most cases it can be considered an implicit argument within the clause (see also Section 5.1.1). As Figure 7.1 shows, in a dependency parse created by the Stanford parser (Klein and Manning, 2002), "John" is marked as the subject of both "entered" and "noticed."

(1) John **entered** the room
and **noticed** the letter on the table.

We look for nominal and clausal subjects, representing them by a reference to their head token. For each main verb, we check whether it has a dependent using one of the relations `nsubj`, `nsubjpass`, `csubj`, `csubjpass` and `xsubj` from the set of Stanford's typed dependency relations (de Marneffe and Manning, 2008b). In addition, we apply several parser-version specific rules that correctly identify the subjects of copula relations and implement fallback strategies for frequently occurring imprecise parses.

7.1.3 Part-of-speech tags (pos)

This set of features comprises one entry per POS tag. For each of the POS tags as used in the Penn Treebank (Santorini, 1990; Marcus et al., 1993) and automatically assigned by the Stanford POS tagger (Toutanova and Manning, 2000; Toutanova et al., 2003), we indicate whether it appears in the segment.

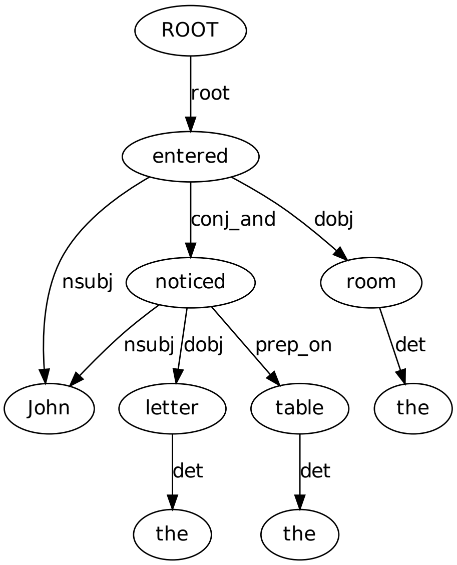

Figure 7.1: Stanford dependency parse for example (1): "John entered the room and noticed the letter on the table."

7.1.4 Brown clusters (bc)

The only existing previous work on labeling situation entities with their type (Palmer et al., 2007, see Section 3.1.4) uses words as features. These simple features work well on their small data set from a limited domain, but the approach quickly becomes impractical with increasing corpus size and variety as in our setting. Besides, the word features overfit the domain (Friedrich et al., 2016).

We instead turn to distributional information in the form of Brown clusters (Brown et al., 1992), which can be learned from raw text and and represent word classes in a hierarchical way. An algorithm for assigning these classes to words was originally developed in the context of n-gram language modeling with the aim of overcoming sparsity issues, the intuition being that predictions for unseen n-gram histories can be improved by modeling similarity of histories. The algorithm seeks to assign words to classes resulting in a partition that maximizes the average mutual information of the words in the classes. For vocabulary size V, the algorithm starts with V clusters and in each step merges the two clusters for which the loss in average mutual information is the least. After $V - C$ steps, a clustering with C clusters remains. The history of the merging steps corresponds to the clustering's hierarchy.

We use existing, freely-available clusters trained on news data by Turian et al. (2010) using the implementation by Liang (2005).[2] We replace each word in a clause with its Brown cluster identifier. Clusterings with 320 and 1000 Brown clusters work best for our task, i.e, we use 1320 numeric Brown cluster features. For each cluster, we count how often a

[2] Precomputed Brown clusters trained on the Reuters Corpus Vol. 1 (`http://trec.nist.gov/data/reuters/reuters.html`) are available from `http://metaoptimize.com/projects/wordreprs`.

word in the clause was assigned to it (most often 0). We additionally experimented with using the Brown cluster identifier of the main verb's lemma or the main referent's head's lemma as a feature, which did not result in improvements.

7.1.5 Main verb (mv)

This set of features captures syntactic-semantic properties of the situation entity's main verb. Features based on WordNet (Fellbaum, 1998) use the most frequent sense of the lemma, including also the sense of the corresponding synset's direct hypernym. Tense, voice and grammatical aspect information is extracted from sequences of POS tags using a set of rules (Loaiciga et al., 2014). We also capture the lemma of the verb's object and whether there is an auxiliary or a particle.

In the next section, we describe a set of type-based linguistic indicator features. In our experiments on automatic classification of situation entity types (Section 8.5), the set abbreviated as **mv** includes those linguistic indicator features for the main verb's lemma. For reasons of readability, we describe the linguistic indicator features in a separate section as they are also used separately in some of our experiments.

7.1.6 Linguistic indicators (lingInd)

This set of corpus-based features is a reimplementation of the linguistic indicators from Siegel and McKeown (2000), who show that (some of) these features correlate with either stative or dynamic verb types (see Section 3.1). Siegel and McKeown learn the feature values from about 100,000 clauses taken from medical discharge summaries parsed automatically with the English slot grammar parser (McCord, 1990). In our replication, we aim at both covering a larger set of verbs and domains, and to use freely available software. Thus, we use the list of indicators and the lists of adverbials as provided by Siegel (1998b), but using a different parsed background corpus. We parse the AFE and XIE sections of GigaWord (Graff et al., 2003) with the Stanford dependency parser, using all documents tagged as "story." For each verb type, we obtain a normalized count showing how often it occurs with each of the indicators in Table 7.2, resulting in one value per feature per verb. For example, for the verb *fill*, the value of the feature `temporal-adverb` is 0.0085, meaning that 0.85% of the occurrences of *fill* in the corpus are modified by one of the temporal adverbs on the list compiled by Siegel (1998a). Tense, Progressive, Perfect and voice are extracted using a set of rules following Loaiciga et al. (2014), which make use of the Penn TreeBank part-of-speech tags (Marcus et al., 1993). For example, the verb tag sequence VBD VBG, i.e., a verb in past tense followed by a gerund or present participle form (e.g., "was running"), is marked as simple past tense and Progressive aspect.

7.1.7 Main referent (mr)

In most cases, the main referent corresponds to a noun phrase (NP). The set of **mr** features thus focuses on describing such. These NP-level features (see also Friedrich and

feature	example
frequency	-
past	*said*
perfect	*had won*
progressive	*is winning*
negated	*not/never*
particle	*up/in/...*
no subject	-
continuous adverb	*continually*, *endlessly*
evaluation adverb	*better*, *horribly*
manner adverb	*furiously*, *patiently*
temporal adverb	*again*, *finally*
in-PP	*in an hour*
for-PP	*for an hour*

Table 7.2: LingInd feature set and examples for lexical items associated with each indicator (Siegel and McKeown, 2000).

Pinkal, 2015a) are inspired by Reiter and Frank (2010), who aim to classify the genericity of NPs. The features include the lemma of the NP's head, its determiner type, noun type, number, WordNet sense and supersense, dependency relations linked to this token, person, countability, and whether it is a bare plural. The countability features are taken from Celex 2 (Baayen et al., 1996) for each lemma and take on the values *count*, *uncount* and *ambig*. As the Celex 2 database is not publicly available, we also extract a list with countability information for lemmas from WebCelex.[3] This feature set is integrated in the publicly available version of our software. The values are *Y* and *N* in this case, and performance for the various classification tasks was slightly worse than when using Celex 2. We thus here present results using Celex 2.

7.1.8 Clause (cl)

These features describe properties at the clause-level, capturing both grammatical phenomena such as word order and lexical phenomena such as presence of particular adverbials or prepositional phrases, as well as semantic information such as modality and negation. Word order, specifically whether the subject occurs before or after the main verb, is an indicator of whether the sentence is a question or not. The feature set also captures whether the clause is a conditional using "if" or "whether." We also use features describing the verb under which the clause's main verb is embedded in a `ccomp` relation, as some verbs embed particular types of situation. For example, the predicate "force" generally embeds verb constellations with dynamic lexical aspectual class as illustrated by example (2). Most of these features are inspired by the work of Palmer et al.

[3] `http://celex.mpi.nl`

(2007) and Reiter and Frank (2010).

(2) (a) John forced Mary to eat the cake.
(b) #John forced Mary to be pretty.

7.2 Models and machine learning methods

We make use of three state-of-the-art supervised machine learning methods; Random Forest classifiers, maximum entropy classifiers and conditional random fields. We here briefly explain each method in general and then explain the various combinations of feature groups and classifiers that we use in our experiments in Chapter 8.

7.2.1 Random Forest classifiers

Random Forests are an ensemble learning method combining several decision trees via bagging (Breiman, 2001). Bagging means that for training each tree of the ensemble, a subset of the training data is sampled with replacement. These classifiers are called "random" as the algorithm makes use of randomized operations for two steps during training. The first one has already been mentioned above; it is the bagging sampling step. Second, random feature selection is applied when deciding how to split a node during the process of expanding the individual decision trees. Finally, the prediction of the Random Forest classifier is computed by taking the majority vote of all trees. Breiman (2001) shows that Random Forests are relatively robust to outliers and noise, and that they achieve in general good accuracy compared to similar algorithms. They can easily combine categorical and numeric features, which makes them a great choice for some of our classification tasks that apply both types of features. In addition, training and prediction are both very fast. We use the Random Forests implementation provided by Weka (Hall et al., 2009), using the default parameter settings.

7.2.2 Maximum entropy classifiers

Maximum entropy classifiers have been widely used in natural language processing (NLP) for decades. Outside the NLP community, maximum entropy classifiers are usually called multi-class logistic regression. They are discriminative classifiers that predict a label y from a set of labels Y for an instance represented by a feature vector x. Maxmimum entropy classifiers make use of feature functions $f_i(x, y)$, which are indicator functions for all combinations of feature values and labels, for example:

```
f = if ($y$ = STATE and $x_j$.perfect=true)
      return 1 else return 0
```

The model decides on the final score by computing a weighted sum of the values returned by the feature functions (see Klinger and Tomanek, 2007):

$$P(y|x) = \frac{1}{Z(x)} exp(\sum_{i=1}^{m} \lambda_i f_i(x, y))$$

Here, λ_i are the weights for the various feature functions. The optimal values for λ_i are computed by optimizing the log-likelihood of the training data; a variety of numerical optimization techniques, e.g., stochastic gradient descent, can be used.
$Z(x)$ is a normalization constant, i.e., the sum of the scores of all possible labels.

$$Z(x) = \sum_{y \in Y} exp(\sum_{i=1}^{m} \lambda_i f_i(x, y))$$

As explained in more detail in the next section, we use a special configuration of the CRF++ toolkit[4] to create maximum entropy models.

7.2.3 Conditional random fields

We use linear chain conditional random fields (Lafferty et al., 2001) to label sequences of mentions or sequences of clauses with regard to their genericity or with regard to their situation entity type. Conditional random fields (CRFs) are well suited for our labeling tasks as they do not assume independence between the features. CRFs predict the conditional probability of label sequence $\vec{y}$ given an observation sequence $\vec{x}$ as follows:

$$P(\vec{y}|\vec{x}) = \frac{1}{Z(\vec{x})} exp(\sum_{j=1}^{n} \sum_{i=1}^{m} \lambda_i f_i(y_{j-1}, y_j, \vec{x}, j))$$

$Z(\vec{x})$ is a normalization constant, the sum over the scores of all possible label sequences for an observation sequence $\vec{x}$ with length n (see also Klinger and Tomanek (2007)). The weights λ_i of the m feature functions are the parameters to be learned. They do not depend on the current position j in the sequence. The feature functions f_i are in general allowed to look at the current label y_j, the previous label y_{j-1} and the entire observation sequence $\vec{x}$. We create a linear chain CRF model using the CRF++ toolkit, using all the default parameters. We use a simple instantiation of a linear chain CRF whose feature functions take two forms, $f_i(y_j, x_j)$ and $f_i(y_{j-1}, y_j)$. The former consists of indicator functions for combinations of labels, e.g., a segment's situation entity type or whether its main referent is generic, and each of the features explained in Section 7.1. They are also called "unigram" functions in CRF++ terminology. Each feature function $f_i(y_j, x_j)$ is an indicator function combining the current label and one of the feature values of the current item, for example:

```
f = if ($y_j$ = EVENT and $x_j$.np.person=3)
     return 1 else return 0
```

The latter type of feature functions $f_i(y_{j-1}, y_j)$, also called "bigram" functions, gets instantiated as indicator functions for each combination of labels, thereby enabling the model to take sequence information into account. Note that while these bigram feature functions are defined over pairs of labels, the label sequence $\vec{y}$ for an observation sequence $\vec{x}$ is optimized over the entire sequence. This means that the choices of labels assigned to non-adjacent clauses *do* influence each other. When using only the former type of feature function, our classifier is equivalent to a maximum entropy model.

[4] https://taku910.github.io/crfpp

Chapter 8

Experimental evaluation

This chapter reports on our computational experiments on automatically labeling clauses with their situation entity types. We break up the problem into three sub-tasks, which correspond to the annotation layers of our corpus as explained in Chapter 5. Our models automatically predict

- whether the lexical aspectual class of a clause's main verb is ***stative*** or ***dynamic*** (Section 8.2);
- whether a clause is ***habitual***, ***episodic*** or ***static*** (Section 8.3);
- and whether the subject (main referent) of a clause is ***generic*** or ***non-generic***, i.e., whether it refers to a kind or not (Section 8.4).

We then combine the methods and features found useful for these subtasks to create our sequence labeling models for **situation entity types**, which we describe in Section 8.5.

8.1 Experimental settings and upper bound

If not otherwise stated, we develop our models using **10-fold cross validation (CV)** on 80% (counted in terms of the number of situation entities) of the MASC and Wikipedia data (a total of 32855 annotated situation entities), keeping the remaining 20% as a held-out test set. Development and test sets each contain distinct sets of documents; the documents of each MASC genre and of Wikipedia are distributed over the folds. Instances from one document are always put in the same fold. This results in slight variations in fold size, but ensures no unfair bias due to very similar instances from the same document.

We report results in terms of macro-average precision, recall and F1-measure (harmonic mean of macro-average precision and macro-average recall), as well as accuracy. We apply McNemar's test with Yates' correction for continuity (McNemar, 1947) with $p < 0.01$ to test significance of differences in accuracy. In the tables in this chapter, we mark numerically-close scores with the same symbols if they are found to be significantly different.

Upper bound: human performance. Labeling clauses with their situation entity types is a non-trivial task even for humans, as there are many borderline cases (see Chapter 6). We compute an upper bound for system performance by iterating over all clauses: for each pair of human annotators, two entries are added to a co-occurrence matrix (similar to a confusion matrix), with each label serving once as "gold standard" and once as the "prediction." From this matrix, we can compute scores in the same manner as for system predictions. Precision and recall scores are symmetric in this case, and accuracy corresponds to observed agreement.

8.2 Automatic prediction of lexical aspectual class

In this set of experiments (see Friedrich and Palmer, 2014a), we describe a new approach to predicting the fundamental aspectual class of verbs in context, i.e., whether a verb is used in a ***stative*** or in a ***dynamic*** sense. The corresponding linguistic theory has been explained in Section 2.1.1 and the data we work with here was annotated according to the guidelines explained in Section 5.1.2. We identify two challenging cases of this problem: when the verb is unseen in training data, and when the verb is ambiguous for aspectual class. A semi-supervised approach using linguistically-motivated features and a novel set of distributional features based on representative verb types allows us to predict lexical aspectual class accurately, even for unseen verbs.

While most verbs have one predominant interpretation, others are more flexible for aspectual class and can occur as either ***stative*** (1) or ***dynamic*** (2) depending on the context. There are also cases that allow for both readings, such as (3), taken from the Brown corpus.

(1) The liquid **fills** the container. (***stative***)

(2) The pool slowly **filled** with water. (***dynamic***)

(3) Your soul was made to be **filled** with God Himself. (***both***)

Cases like (3) that did not result in a majority vote for either ***stative*** or ***dynamic*** were assigned the label ***both*** for the purpose of these experiments. The third label ***both*** does not imply that there is in fact a third class on the level of ***stative*** and ***dynamic***, but rather that two interpretations are available for the sentence, of which usually one will be chosen by a reader.

The main previous experiment that our work builds on is the one by Siegel and McKeown (2000) as explained in detail in Section 3.1.2. In contrast to Siegel and McKeown, we do not conduct the task of predicting aspectual class solely at the type level, as such an approach ignores the minority class of ambiguous verbs. Instead we predict the aspectual class of verbs in the context of their arguments and modifiers. We show that this method works better than using only type-based features, especially for verbs with ambiguous aspectual class. In addition, we show that type-based features, including novel distributional features based on representative verbs, accurately predict predominant aspectual class for unseen verb types. Our approach also differs from prior work in that we treat the problem as a three-way classification task, predicting ***dynamic***, ***stative*** or ***both*** as the aspectual class of a verb in context.

	complete data set		w/o *have/be*	
genre	**clauses**	κ	**clauses**	κ
jokes	3462	0.85	2660	0.77
letters	1848	0.71	1444	0.62
news	2565	0.79	2075	0.69
all	7875	0.80	6179	0.70

Table 8.1: Asp-MASC: Cohen's observed unweighted κ.

		Annotator 2		
		dynamic	***stative***	***cannot decide***
Annotator 1	***dynamic***	4464	164	9
	stative	434	1056	29
	cannot decide	5	0	0

Table 8.2: Asp-MASC: confusion matrix for two annotators, without *have/be* clauses.

8.2.1 Data

Verb type seed sets. Using the **LCS Verb Database** (Dorr, 2001), we identify sets of verb types whose senses are only ***stative*** (188 verbs, e.g. *belong*, *cost*, or *possess*), only ***dynamic*** (3760 verbs, e.g. *alter*, *knock*, *resign*), or mixed (215 verbs, e.g. *fill, stand, take*), following the procedure described by Dorr and Olsen (1997, see also Section 3.1.1).

Asp-MASC. The Asp-MASC corpus consists of 7875 clauses from the letters, news and jokes sections of MASC (Ide et al., 2010), each labeled by two annotators for the aspectual class of the main verb.[1] We use 6161 clauses for the classification task, omitting clauses with *have* or *be* as the main verb and those where no main verb could be identified due to parsing errors (*none*). Table 8.1 shows inter-annotator agreement; Table 8.2 shows the confusion matrix for the two annotators. Our two annotators exhibit different preferences on the 598 cases where they disagree between ***dynamic*** and ***stative***. We observe higher agreement in the jokes and news subcorpora than for letters; texts in the letters subcorpora are largely argumentative and thus have a different rhetorical style than the more straightforward narratives and reports found in jokes. Overall, we find substantial agreement.

The data for our experiments uses the label ***dynamic*** or ***stative*** whenever annotators agree, and ***both*** whenever they disagree or when at least one annotator marked the clause as ***cannot decide***, assuming that both readings are possible in such cases.

Asp-Ambig (Brown). In order to facilitate a first study on ambiguous verbs, we select 20 frequent verbs from the 'mixed' verb types in the LCS seed verb lists and for each

[1] Annotations of a third annotator had not yet been added at the time of these experiments.

		Annotator 2		
		dynamic	*stative*	*cannot decide*
Annotator 1	*dynamic*	1444	201	54
	stative	168	697	20
	cannot decide	44	31	8

Table 8.3: Asp-Ambig: confusion matrix for two annotators. Cohen's κ is 0.6.

annotate 138 sentences. Sentences are extracted randomly from the Brown corpus, such that the distribution of ***stative/dynamic*** usages is expected to be natural. We present entire sentences to the annotators who mark the aspectual class of the verb in question as highlighted in the sentence. Again, we discard instances with parsing problems. This results in 2667 instances. κ is 0.6, the confusion matrix is shown in Table 8.3. Details are listed in Table 8.8.

8.2.2 Computational model

For predicting the aspectual class of verbs in context as ***stative***, ***dynamic*** or ***both***, we assume a supervised learning setting and train a Random Forest classifier (see Section 7.2.1) using the following feature sets:

- **Linguistic indicator features (LingInd)**: This feature set contains the type-based features representing corpus-based usage patterns of verb types as described in Section 7.1.6.

- **Distributional features (Dist)**: Using an existing large distributional model (Thater et al., 2011) estimated over the set of Gigaword documents marked as stories, for each verb type, we build a syntactically informed vector representing the contexts in which the verb occurs. We compute three numeric feature values per verb type, which correspond to the average cosine similarities with the verb types in each of the three seed sets for ***stative***, ***dynamic*** and mixed verbs extracted from LCS.

- **Instance-based features (Inst)**: This feature set is a subset of the main verb features explained in Section 7.1.5. In contrast to the above described type-based features, these features do not rely on a background corpus, but are extracted from the clause being classified. The subset includes the part-of-speech tag of the verb, its tense and voice, and whether it occurs in the Progressive or Perfect. For features encoding grammatical dependents, we focus on a subset of grammatical relations. The feature value is either the WordNet lexical filename (e.g. *noun.person*) of the given relation's argument or its POS tag, if the former is not available. We simply use the most frequent sense for the dependent's lemma. We also include features that indicate, if there are any, the particle of the verb and its prepositional dependents. For the sentence *A little girl had just finished her first week of school*, the instance-based feature values would include `tense`:*past*, `subj`:*noun.person*, `dobj`:*noun.time* or `particle`:*none*.

8.2.3 Results

The experiments presented in this section aim to evaluate the effectiveness of the feature sets described in the previous section, focusing on the challenging cases of verb types unseen in the training data and highly ambiguous verbs. The feature **Lemma** indicates that the verb's lemma is used as an additional feature.

Experiment 1: Seen verbs

The setting of our first experiment follows Siegel and McKeown (2000), using 10-fold cross validation with occurrences of all verbs in Asp-MASC distributed evenly over the folds. Table 8.4 shows the corresponding results. No feature combination significantly outperforms the baseline of simply memorizing the most frequent class of a verb type in the respective training folds. The instance-based features do not work well on their own in this setting, indicating that the task of automatically classifying lexical aspectual class cannot be solved from only looking at a verb's context, but also requires lexical information about the verb type.

Features	**Accuracy (%)**
Baseline (Lemma)	83.6
LingInd	83.8
Inst	70.8
Inst+Lemma	83.7
Dist	83.4
LingInd+Inst+Dist+Lemma	84.1

Table 8.4: Experiment 1: Seen verbs, using **Asp-MASC**.

Experiment 2: Unseen verbs

This experiment shows a successful case of semi-supervised learning: while type-based feature values can be estimated from large corpora in an unsupervised way, some labeled training data is necessary to learn their best combination. This experiment specifically examines performance on verbs not seen in labeled training data. We use 10-fold cross validation on Asp-MASC but ensure that all occurrences of a verb type appear in the same fold: verb types in each test fold have *not* been seen in the respective training data, ruling out the Lemma feature. A maximum entropy classifier works better here (as we use only numeric features). We here implement the classifier using Weka's (Hall et al., 2009) logistic regression. We present results in Table 8.5. The baseline labels everything with the most frequent class in the training set (***dynamic***). Both the LingInd and Dist features generalize across verb types, and their combination works best.

Features	Accuracy (%)
Baseline	72.5
Dist	78.3*
LingInd	80.4*
LingInd+Dist	81.9†

Table 8.5: **Experiment 2**: Unseen verb types, **Asp-MASC**. *Significantly different from baseline. †Significantly different from results for LingInd.

Data	Features	Accuracy (%)
one-label verbs (1966 inst.)	Baseline	92.8
	LingInd	92.8
	Dist	92.6
	Inst+Lemma	91.4*
	LingInd+Inst+Lemma	92.4
multi-label verbs (4195 inst.)	Baseline	78.9
	LingInd	79.0
	Dist	79.0
	Inst	67.4*
	Inst+Lemma	79.9
	LingInd+Inst+Lemma	80.9*
	LingInd+Inst+Lemma+Dist	80.2*

Table 8.6: **Experiment 3**: One- vs. Multi-label verbs, **Asp-MASC**. Baseline as in Table 8.4. *Indicates that result is significantly different from the respective baseline.

Experiment 3: one- vs. multi-label verbs

For this experiment, we compute results separately for one-label verbs (those for which all instances in Asp-MASC have the same label) and for multi-label verbs (instances have differing labels in Asp-MASC). We expect one-label verbs to have a strong predominant aspectual class, and multi-label verbs to be more flexible. Otherwise, the experimental setup is as in experiment 1. Results appear in Table 8.6. In each case, the linguistic indicator features again perform on par with the baseline. For multi-label verbs, the feature combination Lemma+LingInd+Inst leads to significant improvement of 2% gain in accuracy over the baseline; Table 8.7 reports detailed class statistics and reveals a gain in F-measure of 3 percentage points over the baseline. To sum up, Inst features are essential for classifying multi-label verbs, and the LingInd features provide some useful prior. These results motivate the need for an instance-based approach.

System	Class	Accuracy	Precision	Recall	F1
baseline	micro-avg.	78.9	75	79	76
LingInd+Inst+Lemma	***dynamic***		84	95	89
	stative		76	69	72
	cannot decide		51	24	33
	micro-avg.	80.9*	78	81	**79**

Table 8.7: Experiment 3: MULTI-LABEL, precision, recall and F1-score, detailed class statistics for the best-performing system from Table 8.6.

Experiment 4: INSTANCE-BASED classification

For verbs with ambiguous aspectual class, type-based classification is not sufficient, as this approach selects a dominant sense for any given verb and then always assigns that. Therefore we propose handling ambiguous verbs separately. As Asp-MASC contains only few instances of each of the ambiguous verbs, we turn to the Asp-Ambig dataset. We perform a Leave-One-Out (LOO) cross validation evaluation, with results reported in Table 8.8. The third column also shows the outcome of using either only the Lemma, only LingInd or only Dist in LOO; all have almost the same outcome as using the majority class, numbers differ only after the decimal point. Using the Inst features alone (not shown in Table 8.8) results in a micro-average accuracy of only 58.1%: these features are only useful when combined with the feature Lemma. For classifying verbs whose most frequent class occurs less than 56% of the time, Lemma+Inst features are essential. Whether or not performance is improved by adding LingInd/Dist features, with their bias towards one aspectual class, depends on the verb type. It is an open research question which verb types should be treated in which way.

8.2.4 Discussion and conclusion

We have described a new, context-aware approach to automatically predicting aspectual class, including a new set of distributional features. Our experiments show that in any setting where labeled training data is available, improvement over the most frequent class baseline can only be reached by integrating instance-based features, though type-based features (LingInd, Dist) may provide useful priors for some verbs and successfully predict predominant aspectual class for unseen verb types. Our results indicate that in order to arrive at a globally well-performing system, a multi-stage approach is needed: such an approach would treat verbs differently according to whether training data is available and whether or not the verb's aspectual class distribution is highly skewed.

The experiments described in this section are a first important step in determining relevant features for classifying the situation entity type of a clause. Lexical aspectual class of a clause's main verb is necessary for distinguishing EVENT from STATES. The finding that linguistic indicator features work well in combination with instance-based features that represent the clause in which a verb occurs are the basis for the experiments described in

Verb	# of inst.	Majority class		Inst+Lemma	Inst+Lemma +LingInd+Dist
feel	128	**96.1**	STAT	93.0	93.8
say	138	**94.9**	DYN	93.5	93.5
make	136	**91.9**	DYN	91.9	91.2
come	133	**88.0**	DYN	87.2	87.2
take	137	**85.4**	DYN	85.4	85.4
meet	130	83.9	DYN	86.2	**87.7**
stand	130	80.0	STAT	79.2	**83.1**
find	137	**74.5**	DYN	69.3	68.8
accept	134	**70.9**	DYN	64.9	65.7
hold	134	**56.0**	BOTH	43.3	49.3
carry	136	55.9	DYN	55.9	**58.1**
look	138	55.8	DYN	72.5	**74.6**
show	133	54.9	DYN	**69.2**	68.4
appear	136	52.2	STAT	**64.7**	61.0
follow	122	51.6	BOTH	**69.7**	65.6
consider	138	50.7	DYN	61.6	**70.3**
cover	123	50.4	STAT	46.3	**54.5**
fill	134	47.8	DYN	**66.4**	62.7
bear	135	47.4	DYN	**70.4**	67.4
allow	135	37.8	DYN	48.9	**51.9**
micro-avg.	2667	66.3		**71.0***	**72.0***

Table 8.8: Experiment 4: INSTANCE-BASED. **Accuracy** (in %) on **Asp-Ambig**. *Differs significantly from the majority class baseline.

the following.

8.3 Automatic recognition of habituality and clausal aspect

The experiments described in this section provide the first fully automatic approach for classifying clauses with respect to their aspectual properties as ***habitual***, ***episodic*** or ***static***. We build on the work by Mathew and Katz (2009, see Section 3.2.1), which addresses only the ***episodic-habitual*** distinction for dynamic verbs, and on our own work classifying verbs as ***stative*** or ***dynamic*** as described in the previous section. Our method combines different sources of information found to be useful for these tasks, (a) syntactic-semantic features reflecting the local context, i.e., each clause itself, and (b) type-based features representing a lexical profile of verb usage for each verb type. The work presented in this section is the first that exhaustively classifies *all* clauses of a text according to their clausal aspect.

8.3.1 Data

In this section, we describe the data sets used in our experiments.[2]

Penn TreeBank (M&K) data set. Mathew and Katz (2009) randomly select sentences for several verbs from the WSJ and Brown corpus sections of the Penn Treebank. They require the verb to be lexically dynamic. Sentences are marked as **habitual** or **episodic**, further details on the annotation guidelines are not specified. Their data set contains 2743 annotated sentences for 239 distinct verb types. Mathew and Katz remove verb types with highly skewed distributions of labels, but their filtered data set is not available. We follow their filtering approach, but we could not replicate their filtering step. Our final data set contains 1230 sentences for 54 distinct verb types. Mathew and Katz (2009) state that their data set comprises 1052 examples for 57 verb stems. We aimed at producing a similar distribution of labels: our data set contains 73.3% episodic cases, M&K's version has 73.1%.

Wikipedia corpus. In our corpus annotated with situation entities as described in Part II of this thesis, each clause is labeled as ***episodic***, ***habitual*** or ***static***. The guidelines are explained in Section 5.1.4; agreement statistics are given in Section 6.3.3. For the experiments presented in this section, we make use of the 10355 clauses from the Wikipedia subcorpus. Table 8.9 shows the distribution of clausal aspect labels in the gold standard, which contains the cases where at least two annotators agreed on the label. We found only 86 cases where all annotators disagree, and manual inspection shows that most of these cases are related to disagreements on the lexical aspectual class that coincide with an attention slip by one of the annotators.

[2] All data sets are freely available from `www.coli.uni-saarland.de/projects/sitent`. We thank Thomas A. Mathew and Graham Katz for allowing us to publish their data set.

Label	# clauses	% clauses
static	6184	59.7
episodic	2114	20.4
habitual	2057	19.9
total	10355	-

Table 8.9: Wikipedia data, distribution of labels for clausal aspect.

8.3.2 Method

In order to investigate in which circumstances the task of predicting a clause's label (***episodic***, ***habitual*** and ***static***) can be addressed jointly, or whether a pipelined approach is better, we apply the following methods. Our Joint model learns the decision boundaries for the three classes jointly, i.e., as a three-way classification task. In addition, we test a Cascaded model, which uses two models learned for the two different subtasks: (a) identifying static clauses and (b) distinguishing episodic and habitual clauses.

First, we train a model to distinguish the ***static*** class from the other two. In this learning step, we simply map all the clauses labeled as ***episodic*** and ***habitual*** to the class ***non-static*** and learn the decision boundary between the two classes ***static*** and ***non-static***. Second, we train a model to distinguish the ***episodic*** from the ***habitual*** class. This model is trained on the subset of examples labeled with either of these two classes.

In the Cascaded model, first, the ***static*** vs. ***non-static*** model is applied. The Cascaded model labels all instances automatically labeled as ***static*** in this first step, and then applies the second model (***episodic*** vs. ***habitual***) on all remaining instances.

We train Random Forest classifiers (see Section 7.2.1) for each step and also for the Joint model, making use of the following two feature sets to describe each clause:

- Context-based features: Table 8.10 shows the syntactic-semantic features, which we call context-based as they are extracted from the context of each verb occurrence that we classify. This feature set comprises the features proposed by Mathew and Katz (2009) and the ones proposed by Friedrich and Palmer (2014a). In addition, we use the features *modal* and *negated*. The features have been extracted as described in Chapter 7. The values of the grammatical dependents' features are the WordNet (Fellbaum, 1998) lexical filename of the dependent's lemma, or, if not available, the dependent's part-of-speech tag. Further details on the features adopted from Mathew and Katz (2009) are given below.

- Type-based features: This feature set consists of the verb-type based linguistic indicator features of Siegel and McKeown (2000). They are explained in Section 7.1.6.

Besides providing a robust performance, Random Forest classifiers can easily deal with both categorical and numeric features. This is relevant as our Context-based features

Feature		Values
verb	tense*†	past, present, infinitive
	pos†	VB, VBG, VBN, ...
	voice†	active, passive
aspect	progressive*†	true, false
	perfect*†	true, false
subject	bare plural*	true, false
	definite*	true, false
	indefinite*	true, false
object	absent*	true, false
	bare plural*	true, false
	definite*	true, false
	indefinite*	true, false
grammatical dependents†		WordNet lexname/POS
sentence	modal	*would, can,...*
	negated	true, false
	conditionals*	presence of clause starting with if/when/whenever
	temporal modifiers*	specific, quantificational, including *used to* and *would* (where no if)
	prepositions*	at / in / on (3 features, true/false)

Table 8.10: Context-based features. Used by: *Mathew and Katz (2009), †Friedrich and Palmer (2014a).

are categorical while the Type-based features are numeric. In our experiments, we will compare the impact of the different feature sets on each subtask and on the Joint model.

Baseline: Mathew and Katz (2009). As a baseline, we also report results for the subset of our Context-based features used by Mathew and Katz (2009) and call this subset M&K. Mathew and Katz (2009) compare several machine learning algorithms on their classification task. They find a J48 decision tree and a Naive Bayes classifier to work best. We replicate their results for the decision tree in Section 8.3.3.

As shown in Table 8.10, we use the features used by this baseline in our system as well. Quantificational adverbs are temporal modifiers such as *always*, *occasionally* or *weekly*.[3] Specific temporal adverbs are, according to a heuristic proposed by Mathew (2009), phrasal children of verbs marked with the part-of-speech tag TMP. Noun phrases with one of the determiners *the, this, that, these, those, each, every, all*, as well as possessives, pronouns, proper names and quantified phrases are definite. NPs with determiners *a, an, many, most, some*, and cases of modifying adjectives without determiners (e.g., *few*) or cardinal numbers (part-of-speech tag CD) are indefinite. Mathew (2009) describes their features in detail.

[3] The complete list of quantificational adverbs used is given by Mathew (2009), page 36.

8.3.3 Experiments and discussion

We now describe our experimental results and discuss them. First, we reproduce the experiments of Mathew and Katz (2009), who use manually created syntactic parses, in a purely automatic setting. The data set and experiments of Mathew and Katz (2009) focus on the ***episodic-habitual*** distinction using a set of sentences selected for a small set of verbs, and their feature design focuses on syntactic properties of the clauses found in this annotated data set. In the further experiments, we turn to the Wikipedia data, which contains annotations for full texts. We expect the Wikipedia data to cover the range of habitual and episodic expressions more fully, and in addition, allows for studying the task of separating static sentences from the other two classes. As we will show, this latter task profits from including features relevant to the stative-dynamic distinction on the lexical level.

We first present experimental results for the two subtasks described in Section 8.3.2. Our Cascaded model first identifies ***static*** clauses, and then classifies the remaining clauses as ***episodic*** or ***habitual***. For reasons of readability, we first report on our experiments for the ***episodic-habitual*** distinction using both the M&K and Wikipedia data sets. Using the Wikipedia data, we then report on the results for the ***stative*** vs. ***non-static*** distinction. Finally, we turn to the full task of the three-way distinction.

Cross validation settings. We report results for 10-fold cross validation (CV) with two different settings: In the Random CV setting, we randomly distribute the instances over the folds, putting all instances of one document into the same fold. In the Unseen verbs CV setting, we simulate the case of not having labeled training data for a particular verb type by putting all instances of one verb type into the same fold.

Experiment 1: M&K data: *episodic* vs. *habitual*

We use Weka's 10-fold stratified cross validation and a J48 decision tree in the experiments reported in this section in order to replicate their experimental setting. Results are shown in Table 8.11. For the sake of completeness, we also show the results as presented in the original paper. F1-scores are computed from P and R as reported in the original paper. Note that their experiments are performed on a different subset of the data and so these numbers are not directly comparable to ours, but as explained above, our subset has a very similar class distribution. Our accuracies based on automatic parses rather than gold standard parses are about 3% lower when using the original feature set (M&K). We conclude that our results are in the expected range. Also, we do not find any significant improvements on this data set when using any other feature sets or combinations thereof (the table shows the results for our Context-based feature set); the M&K feature set designed for this corpus captures its variation well.

We have used a J48 decision tree in this section for comparability with previous work. In all following experiments in this section, we present results using Random Forest classifiers.

System	F1-score			Accuracy
	episodic	*habitual*	*macro*	
majority class*	84.5	0.0	42.2	73.1
M&K*	91.1	70.5	80.8	86.1
M&K	89.6	63.5	76.5	83.8
Context	90.0	64.7	77.3	84.4

Table 8.11: Experiment 1: Results for **episodic** vs. **habitual**, J48 decision tree, data from Mathew and Katz (2009). *Numbers from original paper.

Experiment 2: Wikipedia: *episodic* vs. *habitual*

We study the classification task of distinguishing ***episodic*** and ***habitual*** sentences using the subset of the Wikipedia data having one of these two labels (4171 instances). This task parallels the experiment of Mathew and Katz (2009) described above. We conduct two experiments, once using the Random CV setting and once using the Unseen verbs setting. Table 8.12 shows the results. The distribution of instances is nearly 50:50 in the gold standard (Table 8.9), and the majority classes in the respective training folds differ (this is the reason for the different baseline scores). For reasons of space we do not show the other scores here; macro-average F1-scores have (almost) the same values as accuracy, the F1-scores for episodic and habitual are similar to each other in each case.

Features	Random CV	Unseen verbs
majority class baseline	42.1	46.3
lemma baseline	65.4	46.3
Type-based	68.1	53.9
M&K	82.3	‡81.4
Context-based	*†82.8	‡**83.8**
Context-based + lemma	*84.3	-
Context-based + Type-based	†**85.1**	83.1
Context-based + Type-based + lemma	84.0	-

Table 8.12: Experiment 2: **Wikipedia**: **Accuracy** of ***episodic*** vs. ***habitual***, Random Forest classifier, 4171 instances, 10-fold cross validation, *†‡respective differences statistically significant.

Our findings are as follows: Type-based features outperform the majority class baseline, which means that some verbs have a preference for being used as either ***episodic*** or ***habitual***. The Context-based features work remarkably well. If training data of the same verb type is available, adding the Type-based features or the lemma to the Context-based features results in improvements; this is not the case in the Unseen verbs setting. The latter setting shows that the additional contextual features (compared to the M&K

subset) are important: our corpus indeed covers a broader range of phenomena than the M&K data set.

	RANDOM CV		UNSEEN VERBS	
Features	**F1**	**Acc.**	**F1**	**Acc.**
majority class baseline	37.4	59.7	37.4	59.7‡
M&K	67.5	*69.5	59.2	62.7‡
CONTEXT-BASED	70.3	*71.7	62.8	64.9‡
CONTEXT-BASED + lemma	81.9	†82.8		
TYPE-BASED	78.8	79.3	72.2	73.2‡
CONTEXT-BASED + TYPE-BASED	83.6	†84.1	**78.4**	**79.2**‡
CONTEXT-BASED + TYPE-BASED + lemma	**83.8**	**84.4**		

Table 8.13: Experiment 3: **Wikipedia**: ***static*** vs. ***non-static***. All 10355 instances, 10-fold cross validation.*†‡ differences statistically significant.

Experiment 3: Wikipedia: *static* vs. *non-static*

We evaluate the task of classifying ***stative*** versus ***non-static*** clauses using all 10355 instances of the Wikipedia data set. Any instance labeled ***episodic*** or ***habitual*** receives the label ***non-static*** both in training and testing. Results of this task are shown in Table 8.13. For this subtask, the CONTEXT-BASED features are less informative than the TYPE-BASED features. Again, using lemma information approximates the use of type-based information, but this is not an option in the UNSEEN VERBS setting. A combination of the CONTEXT-BASED and TYPE-BASED features achieves the best results. In Section 8.2, we have found that TYPE-BASED features generalize well across verb types when predicting the aspectual class of verbs in context, the same is true here. They achieve small improvements by adding context-based features. Predicting the lexical aspectual class of the clause's main verb is only part of our classification task, the ***static*** class includes not only lexically stative clauses but also clauses with lexically dynamic verbs that are stativized, e.g., modals, negation or perfect tense. Hence, as expected, in our task, adding the CONTEXT-BASED features results in a considerable performance improvement (5-7% absolute in accuracy). It is worth noting that even for verbs not seen in the training data, high accuracies and F1-scores of almost 80% can be reached.

Experiment 4: Wikipedia: combined task

In this section, we describe our experiments for the three-way classification task of ***static***, ***episodic*** and ***habitual*** clauses, as in a realistic classification setting, a clause may belong to either of these three classes. We investigate whether a pipelined CASCADED approach is better, or whether the JOINT model profits from learning the decision boundaries between all three classes jointly. The results for this task are presented in Table 8.14 and Table 8.15. Both the CONTEXT-BASED and the TYPE-BASED features when used alone improve over the majority class baseline by about 10% in accuracy in the Random CV setting, and only by about 4% in the Unseen verbs setting. In the latter setting, all feature sets when used alone are ineffective for identifying habituals. This indicates that the Context-based features only "pick up" on some type-based information in the Random CV case. The best models for this Joint classification task use both the Context-based and the Type-based feature sets: F1-scores and accuracy increase remarkably. Again, in the Random CV setting, using the lemma results in a large performance gain, though using the Type-based features is beneficial, and, in the Unseen verbs setting, essential.

We apply the Cascaded model as described in Section 8.3.2, training and testing the models for the subtasks in each fold. In the Random CV setting, the accuracy of the Cascaded approach is not significantly better than the one of the Joint approach, though F1-scores for the less frequent **episodic** and **habitual** classes both increase. In the Unseen verbs setting, however, the difference is remarkable: macro-average F1-score increases by almost 5% (absolute) and accuracy increases by 2.2%. Most notably, the F1-score for the habitual class increases from 0.31 to 0.50 (due to an increase in recall). To conclude, the Cascaded approach is favorable as it works more robustly both for verb types seen or unseen in the training data.

Features	**F1-score**				
	stative	***episodic***	***habitual***	*macro*	**Acc.**
majority class baseline	74.8	0	0	24.9	59.7
Joint: M&K	76.6	65.4	26.1	57.5	*67.0
Joint: Context-based	77.5	65.8	36.4	60.5	*68.4
Joint: Context-based + lemma	85.5	75.0	51.6	71.8	†78.0
Joint: Type-based	81.9	52.7	49.7	61.5	69.9
Joint: Context-based + Type-based	86.1	75.8	58.8	73.8	†79.0
+ lemma	86.8	75.0	59.9	74.2	79.6
Cascaded	**86.9**	**76.1**	**62.2**	**75.1**	**79.9**

Table 8.14: Experiment 4: **Wikipedia**: **static** vs. **episodic** vs. **habitual**. Random Cross validation. The Cascaded model uses the best models from Table 8.13 and Table 8.12. *†‡ differences statistically significant.

Feature ablation

Above, we have compared the two major feature groups of Context-based and Type-based features. In addition, we ablate each single feature from the best results for each experiment. For all classification tasks, we found features reflecting tense and grammatical aspect to be most important, both for the Context-based and Type-based features. In general, we observe that no single feature has a big impact on the results, accuracy drops only by at most 1-2%. This shows that our feature set is quite robust and some of the features (e.g., part-of-speech tag of the verb and tense) reflect partially redundant information. However, using only the best features results in a significant performance

Features	**F1-score**				
	stative	***episodic***	***habitual***	*macro*	**Acc.**
majority class baseline	74.8	0.0	0.0	24.9	‡59.7
Joint: M&K	76.3	41.7	0.8	49.0	‡63.8
Joint: Context-based	74.7	57.1	12.0	51.7	⋆63.9
Joint: Type-based	74.9	4.2	2.8	40.7	⋆60.0
Joint: Context-based + Type-based	81.2	69.5	31.3	63.6	**72.1
Cascaded	**82.6**	**72.0**	**50.2**	**68.4**	****74.3**

Table 8.15: Experiment 4: **Wikipedia**: **static** vs. **episodic** vs. **habitual**. Unseen verb types experiment. The Cascaded model uses the best models from Table 8.13 and Table 8.12. *† ‡ ⋆** differences statistically significant.

drop by several percentage points in the various settings, which means that though single features may not have a large impact, overall, the models for this classification task profit from including many diverse features.

For the ***episodic-habitual*** distinction, in the Unseen Verbs setting, the definiteness of the object was an important Context-based feature. In the ***static*** vs. ***non-static*** task, the subject also plays an important role, as well as the Type-based feature for *continuous adverbs*. In the Unseen verbs setting, many Type-based features are important, including those indicating how often the verb type occurs with *adverbs of manner*, *negation* and *in-PPs* in the background corpus. For the **combined** three-way task, we found the main verb's lemma and the direct object to have most impact. Of the Type-based features, the *for-PP*, *present* and *temporal adverbial* were most important. In the Unseen verbs setting, many linguistic indicator features (among others *past*, *progressive*, *negation*) play a greater role, as well as information about the object, subject and tense.

8.3.4 Discussion

In this section, we have presented an approach for classifying the aspect of a clause as ***habitual***, ***episodic*** or ***static***. Clearly, when exhaustively classifying all clauses of a text, the ***static*** class cannot be ignored as was done in previous work; we have shown that we can separate these instances from ***episodic*** and ***habitual*** instances, most of which are lexically dynamic, with high accuracy. Our model for distinguishing ***episodic*** and ***habitual*** sentences integrates a wide range of contextual information and outperforms previous work. Previous work has only addressed the classification of lexical aspectual class and the automatic distinction of episodic and habitual sentences. Our work is the first bringing together two strands of work relevant to classifying clausal aspect, and we have shown that sources of information relevant to these two underlying aspectual distinctions are relevant for our three-way classification task.

We have shown that for distinguishing ***static*** sentences from the other two, Type-based *and* Context-based information is needed; for distinguishing ***episodic*** and ***habitual*** clauses, Context-based features are most important. Our experimental results show that the three-way classification task is most effectively approached by combining both contextual and verb-type based information. Especially for verbs unseen in the training data, we found the CASCADED approach to work better. It is hard for the JOINT approach to identify habitual clauses; while in the CASCADED approach, performance for both steps is high and adds up.

We found the overall performance of this task to be about 80% accuracy, and 75% macro-average F1-score. These scores suggest that this method may be usable as a preprocessing step for further temporal processing.

Our models do not yet take discourse information into account. Consider example (4) by Mathew and Katz (2009): The second sentence is habitual, but the only indicator for this is sentence-external.

(4) John rarely ate fruit. He just ate oranges. (***habitual***)

In some preliminary experiments, we tried to leverage the discourse context of a clause for its classification by means of incorporating the gold standard label of the previous clause as a feature. This did not result in significant performance improvements. However, further experiments trying to incorporate discourse information are necessary, and, due to our new corpus of fully annotated texts, now possible.

8.4 Discourse-sensitive automatic identification of generic expressions

In this section, we focus on automatically recognizing genericity on two levels: (a) we call *clauses* generic if they provide a general characterization of entities of a certain kind, and (b) we call *NP-level mentions* generic if they refer to kinds or arbitrary members of a class. In terms of our annotation guidelines as explained in Chapter 5, the first task corresponds to recognizing whether a situation entity is a GENERIC SENTENCE; the latter task corresponds to recognizing whether the subject of a clause, i.e., a situation entity's main referent, is ***generic*** or not.

Although genericity on the clause- and NP-level are strongly interrelated, the concepts do not always coincide. As example (5) shows, sentences describing episodic events can have a generic NP as their subject. Note that references to species are kind-referring / generic on the NP level (following Krifka et al. (1995), see p. 65).

(5) In September 2013 <u>the blobfish</u> was voted the "World's Ugliest Animal". (subject ***generic***, clause ***non-generic***)

Genericity often cannot be annotated without paying attention to the wider discourse context. Clearly, coreference information is needed for the genericity classification of pronouns. Often, even genericity of full NPs or entire clauses cannot be decided in isolation, as illustrated by example (6). Sentence (b) could be part of a particular narrative about a tree, or it could be a generic statement. Only the context given by (a) clarifies that (b) indeed makes reference to any year's new twigs and is to be interpreted as generic.

(6) (a) Sugar maples also have a tendency to color unevenly in fall. (***generic***)
(b) The recent year's growth twigs are green and turn dark brown. (***generic***)

In computational linguistics, most research on detecting genericity has been done in relation to the ACE corpora (Mitchell et al., 2003; Walker et al., 2006), focusing on assigning genericity labels to noun phrases (Suh et al., 2006; Reiter and Frank, 2010). We compare to the method of Reiter and Frank (2010), described in detail in Section 3.2, as a highly-competitive baseline.

Our work is based on these approaches, most notably on the work of Reiter and Frank (2010). We present a *discourse-sensitive genericity labeler*. Technically, we use conditional random fields as a sequence labeling method (Section 7.2.3). We train and evaluate our method on the Wikipedia dataset and the ACE corpora, evaluating both the tasks of predicting NP genericity and the task of predicting clause-level genericity.

The ACE corpora are only annotated for NP-level genericity. In our corpus, from which the Wikipedia data set is taken, each clause is manually annotated with the following information (for more details on the annotation scheme, see Friedrich et al. (2015b) and Chapter 5):

- **Task NP**: whether or not the *subject* NP of the clause refers to a class or kind (***generic*** vs. ***non-generic***); this corresponds to our labels for the genericity of a clause's main referent.
- **Task Cl**: whether the *clause* is ***generic***, defined as a clause that makes a characterizing statement about a class or kind, or ***non-generic***; this corresponds to the task of distinguishing Generic Sentence from the other situation entity types.
- **Task Cl+NP**: using the information from Task NP and Task Cl above, we automatically derive the following classification for each *clause* (compare to the explanation of example (5)).
    - **GEN_gen**: generic clause, subject is ***generic*** by definition (*The lion is a predatory cat*); these correspond to Generic Sentences;
    - **NON-GEN_non-gen**: non-generic clause with a ***non-generic*** subject (*Simba roared*); these are all situation segments that are not marked as having a ***generic*** main referent;
    - **NON-GEN_gen**: episodic clause with a generic subject (*Dinosaurs died out*); these correspond to Events with ***generic*** main referents;
    - **GEN_non-gen** does not exist by definition.

For the Wikipedia data, inter-annotator agreement measured as Fleiss' κ (Fleiss, 1971) on the segments labeled by all three annotators is 0.70, 0.73 and 0.69 for Task NP, Task Cl and Task Cl+NP respectively, indicating substantial agreement (Landis and Koch, 1977).

8.4.1 Method and baseline

The system for identifying generic NPs by Reiter and Frank (2010), henceforth R&F, makes use of the English ParGram LFG grammar for the XLE parser (Butt et al., 2002). As this grammar is not publicly available, we implement a similar system using exclusively publicly available resources as described in Chapter 7. The full set of features used in our experiments is shown in Table 8.16. We could not reimplement several tense- and aspect-related ParGram-specific features. In order to compensate for this, we add an additional feature (tense) with finer-grained tense and voice information. Other additional features did not improve performance, which shows that R&F's set of features captures the syntactic-semantic information relevant to genericity classification quite well. Therefore, we use this feature set also for the sequence labeling model. Using the same feature set allows us to attribute any performance gain to the context-awareness of our model rather than the features.

Baseline: R&F. Reiter and Frank (2010) train a Bayesian network using Weka (Hall et al., 2009). The decisions of this classifier are local to each clause. They report the performance of their system on the ACE-2 corpus: Table 8.17 shows that the performance of our re-implemented feature set is comparable to the system of R&F.[4] In all other other tables, "BayesNet R&F" refers to our re-implemented system.

R&F present the "Person baseline" as a simple informed baseline (see Table 8.17). We trained a J48 decision tree on this feature alone, which confirmed that only second-person mentions (the generic "you") are classified as generic, while all other mentions are classified as non-generic.

CRF models. We train two different types of models for each task, representing each clause, or the subject of each clause, using the features shown in Table 8.16.

- **CRF-bigram**: This model is a conditional random field (CRF) including bigram features which represent transition probabilities between the labels occurring in a sequence. It hence is able to take discourse information into account.
- **CRF-unigram**: This model uses exactly the same implementation but excluding the bigram features. It classifies each clause or NP separately – though taking the local clause into consideration for subject NPs – and is thus equivalent to a maximum entropy model.[5]

Comparing the CRF-bigram and the CRF-unigram models enables us to estimate the amount of impact that the bigram contribute. Simply comparing the R&F's model is not sufficient for this purpose because, as we will see below, the CRF-unigram model itself strongly outperforms their Bayesian network. This is simply due to our discriminative formulation of the problem using maximum entropy.

[4] Table 6 in Reiter and Frank's paper contains some typographical errors here. We thank Nils Reiter for making available his ARFF files, so we can provide this updated version.

[5] We call this model CRF-unigram here for consistency with the related publication (Friedrich and Pinkal, 2015a).

NP-based Features	**Values**
number	sg, pl
person	1, 2, 3
countability	from *Celex*: count/uncount/ambig
noun type	common, proper, pronoun
determiner type	def, indef, demon
part-of-speech	POS of head
bare plural	true, false
WordNet granularity	number of edges to top node
WordNet sense $[0-2]$	WordNet senses (head+hypernyms)
WordNet senseTop	top sense in hypernym hierarchy
WordNet lexical filename	person, artifact, event, ...
Clause-based Features	**Values**
dependency $[0-4]$	dependency relation between head and governor etc.
tense	tense, aspect and voice information, e.g. *pres_perf_active*
coarseTense	pres, past, fut
progressive	true, false
perfective	true, false
passive	true, false
temporal modifier	true, false
number of modifiers	numeric
part-of-speech	POS of head
predicate	lemma of head
adjunct-degree	positive, comparative, superlative
adjunct-pred	lemma of adverbial clauses' head

Table 8.16: Features for genericity classification adapted from Reiter and Frank (2010).

System	*generic*			*non-generic*			macro-avg			
	P	R	F1	P	R	F1	P	R	F1	Acc.
majority class baseline	0.0	0.0	0.0	86.8	100	92.9	43.4	50.0	46.5	86.8
person baseline (R&F)	60.4	10.2	17.5	87.9	99.0	93.1	74.2	54.6	62.9	87.2
R&F (BayesNet)	37.7	72.0	49.5	95.0	81.9	88.0	66.4	76.9	71.3	80.6
Reimpl. (BayesNet)	38.1	67.7	48.8	94.4	83.3	88.5	66.3	75.5	70.6	81.2

Table 8.17: Results of **reimplemented baseline** on ACE-2 (original, unbalanced data set), 40106 instances (annotated noun phrases). Weka's stratified 10-fold cross validation, using all features.

8.4.2 Experiments on ACE data

We here present results for experiments using the ACE-2 (Mitchell et al., 2003) and ACE-2005 (Walker et al., 2006) corpora, which have also been used by Reiter and Frank (2010). Details and criticism regarding the related annotation schemes have been presented in Section 3.2.

On the ACE corpora, we only conduct Task NP because there are no labels corresponding to Task Cl or Task Cl+NP. From ACE-2005, we use the newswire and broadcast news subsections.[6] Due to low frequency, we omit instances of NEG in our experiments, and apply a three-way classification task (GEN, SPC, USP). We present results for all remaining 40106 mentions and for the subset of 18029 subject mentions, each time using 10-fold CV.

Both on ACE-2 (see Table 8.18) and on ACE-2005 (see Table 8.19), the CRF outperforms the method of Reiter and Frank (2010) in terms of accuracy, and has a higher F1-score. We give results also for subjects only as this parallels the setting of the Wikipedia experiments (reasons for the restriction to subjects were given in Section 3.2.2). For subjects, the majority class SPC is less frequent (compare the accuracies of the two majority class baselines); only 7% of the subjects are marked as GEN, the rest are labeled as USP. The bigram model does not outperform the unigram model, but our oracle experiments show that context information is indeed useful: accuracy increases significantly and F1 increases considerably, especially for subjects.

We identify two reasons for the fact that when evaluating on the ACE corpora, oracle information is needed to show the benefit of using bigram feature functions: (a) The frequency of **GEN** mentions in the ACE corpora is low – news contains only little generic information, so the context information is harder to leverage. (b) The ACE annotation guidelines contain some vagueness (see Section 3.2.2); this makes it harder for an automatic system to learn about regularities.

[6] The rest of the data comprises broadcast conversation, weblog and forum texts as well as transcribed conversational telephone, and would require specialized preprocessing.

System	***generic***	***non-generic***	**macro-avg**			
	F1	F1	P	R	F1	**Accuracy**
majority class	0.0	92.9	43.4	50.0	46.5	86.8
BayesNet (R&F)	47.4	87.9	65.5	**74.6**	69.8	80.4
CRF (unigrams)	49.1	93.5	75.5	68.7	71.3	88.5*
CRF (bigrams)	**51.0**	**93.7**	**76.5**	69.8	**72.4**	**88.9**
CRF (bigram, gold)	***57.6***	***94.4***	***79.8***	*73.4*	***76.0***	***90.1****

Table 8.18: Results on **ACE-2** for **Task NP**, 10-fold CV, folds contain complete documents. *Difference statistically significant.

System	**macro-avg**			
	P	R	F1	**Accuracy**
all 18029 annotated mentions				
Majority class	27.0	33.3	29.9	81.1
BayesNet (R&F)	50.8	57.2	53.8	74.5
CRF (unigram)	**61.6**	51.8	**55.1**	**83.2***
CRF (bigram)	60.6	51.7	54.8	83.0
CRF (bigram, gold)	***63.9***	***54.9***	***58.2***	***83.9****
5670 subject mentions				
Majority class	25.0	33.3	28.6	75.1
BayesNet (R&F)	51.5	**53.9**	52.7	72.5
CRF (unigram)	58.0	51.3	53.6	77.7*
CRF (bigram)	58.3	51.3	53.7	77.8
CRF (bigram, gold)	***62.4***	***56.1***	***58.6***	***79.6****

Table 8.19: Results on **ACE-2005** (bn+nw), **Task NP**, 10-fold CV, 3 classes: SPC, GEN, USP. *Difference statistically significant.

8.4.3 Experiments on Wikipedia data

In order to study generics in a genre other than news (as in ACE), we turn to an encyclopedia, in which we expect many generics. We use the Wikipedia part of the data described in Part II of this thesis, with the mapping given in the introduction to this section. For the experiments on the Wikipedia data, we use leave-one-document-out CV, i.e., we train on 101 of the 102 documents and test on the remaining document in each fold. The total number of clauses is 10355. We first discuss the results of our experiments in terms of identifying generic NPs or clauses. Then we present some additional experiments testing the influence of the different feature classes and of other discourse-related information.

All tasks, Wikipedia. The observations described in this paragraph are the same for all three prediction tasks on Wikipedia. As Table 8.20, Table 8.21 and Table 8.22 show, our

CRF models outperform the baseline system of R&F by a large margin both in terms of accuracy and F1-score on the Wikipedia corpus. In Task NP and Task Cl, precision and recall are quite balanced (not shown in tables). The performance of the bigram model is significantly better than that of the unigram model, increasing accuracy by about 3%, at the same time increasing F1. In an oracle experiment, we use the previous gold label instead of the predicted one for $f_i(y_{j-1}, y_j)$, and scores increase by up to 6.6% compared to the unigram model. These results provide strong empirical evidence for our hypothesis that using context information is useful for identifying the genericity of NPs or clauses.

Task Cl+NP, Wikipedia. In Task Cl+NP (see Table 8.22), only about 6% of the instances have the gold label **NON-GEN_gen** (i.e., a non-generic sentence with a generic subject), the other instances are distributed roughly evenly between the other two labels. The difficulty of Task Cl+NP thus consists in identifying this infrequent case. The three-way CRF outperforms the two-step approach both in terms of accuracy and macro-average F1-score. The precision-recall tradeoff differs: for the **NON-GEN_gen** class, P and R of the CRF are 55.2% and 24.5% and those of the two-step-approach are 23.8% and 35.9%. The two-step approach labels more instances as **NON-GEN_gen** but does so in a less precise way. While the performance of our model leaves room for improvement on Task Cl+NP, especially with regard to the class **NON-GEN_gen**, it is worth noting that the computational model captures something about the nature of this latter class; its instances *do* look different in the feature space. The context-aware CRF using three labels performs best.

System	***generic***	***non-generic***	**macro-avg.**	
	F1	F1	F1	**Accuracy**
majority class	71.9	0.0	35.9	56.1
BayesNet (R&F)	72.6	70.8	72.3	71.7
CRF (unigram)	79.3	72.6	75.9	76.4*
CRF (bigram)	**81.3**	**76.3**	**78.8**	**79.1***
- only clause features	79.2	71.6	75.5	76.0
- only NP features	76.8	70.8	73.8	74.1
CRF (bigram, gold)	***85.0***	***80.4***	***82.7***	***83.0***

Table 8.20: Results on Wikipedia data for **Task NP** (genericity of subject). *†Difference statistically significant.

Feature set ablation. In this ablation test, shown in Table 8.20 and in Table 8.22, our best model (CRF bigram) uses either the set of clause-based or the set of NP-based features at a time. Clause-based features are more important than the NP-based features for all three classification tasks. An interesting observation is that the NP features alone are not able to separate the infrequent class **NON-GEN_gen** from the other two at all, the F1-score of 2.5 shows that almost all instances of this class were labeled as one of the other two classes. In sum, this shows that whether an NP is interpreted as generic or not strongly depends on how it is used in the clause.

System	*generic* F1	*non-generic* F1	macro-avg. F1	Accuracy
majority class	60.3	3.7	35.1	43.7
BayesNet (R&F)	72.4	74.6	73.7	73.5
CRF (unigram)	77.9	77.0	77.4	77.4†
CRF (bigram)	**80.8**	**80.6**	**80.7**	**80.7**†
- only clause features	79.3	78.3	78.8	78.8
- only NP features	70.7	72.6	71.8	71.7
CRF (bigram, gold)	***82.9***	***82.6***	***82.8***	***82.8***

Table 8.21: Results on **WikiGenerics** for **Task Cl** (clause-level genericity). *†Difference statistically significant.

System	GEN	NON-GEN	NON-GEN				
	gen	non-gen	gen		macro-avg		
	F1	F1	F1	P	R	F1	Accuracy
majority class	67.1	0.0	0.0	16.8	33.3	22.4	50.4
BayesNet (R&F)	69.1	69.1	26.1	54.5	58.4	56.4	65.2
CRF (unigram)	78.5	72.6	**35.4**	67.2	60.0	63.4	74.0*
CRF (bigram)	**81.3**	**76.9**	33.4	**70.3**	61.8	**65.8**	**77.4***
- two-step	80.8	75.8	28.6	61.5	**62.3**	61.9	73.4
- only clause feat.	79.4	72.6	25.3	67.0	57.2	61.8	74.3
- only NP feat.	72.9	71.4	2.5	53.0	49.9	51.4	70.0
CRF (bigram, gold)	***84.0***	***80.6***	***39.1***	***72.8***	***65.7***	***69.0***	***80.6***

Table 8.22: Results on Wikipedia for **Task Cl+NP** (genericity of clause, three-way). *Difference statistically significant.

Higher-order Markov models. Another research question is whether models incorporating not only the previous label, but more preceding labels would perform even better. We turn to the Mallet toolkit McCallum (2002), whose CRF implementation allows for using higher-order models.[7] For example, an order-2 model considers the two previous labels. We use L1-regularization during training. Figure 8.1 shows that the optimum is reached for order-1 (bigram) models for each of the classification tasks for accuracy, the same tendencies were observed for F1-score (not shown). It seems sufficient to use bigram feature functions; note that as explained in Section 7.2.3, the bigram model does not mean that only adjacent clauses influence each other – context is actually wider.

[7] The CRF++ toolkit, which we use in all other experiments, does not allow for higher-order models. We use CRF++ in the main experiments as it comes with a concise documentation; this helps to make our experiments easily replicable.

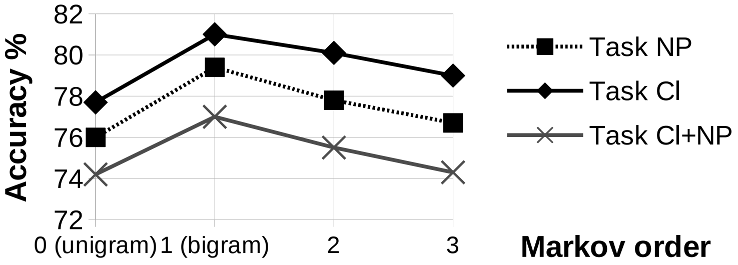

Figure 8.1: Labeling results for CRF models of various orders on Wikipedia corpus.

Using coreference information. In our approximately balanced Wikipedia corpus, 54% of all pronouns are marked as generic and 46% are marked as non-generic, which shows that there is no preference for pronouns to occur with either class. Some of the features (countability, noun type, determiner type, bare plural, and the WordNet related features) are not informative when applied to personal or relative pronouns. Sometimes, it is not even possible to determine number referring to the antecedent (e.g., in the case of the relative pronoun "who"). We conduct the following experiment: we automatically resolve coreference using the Stanford coreference resolution system (Raghunathan et al., 2010). We replace the NP features of each pronominal instance with the features of the first link of the coreference chain. We did not obtain a significant performance gain. One reason is that this change of features only applies to about 13% of the data. We observe that any positive changes in the classification go along with some negative changes which were often due to coreference resolution errors. One difficult step in manually annotating, and hence also in automatically resolving coreference is to determine whether a NP is generic or not (Nedoluzhko, 2013). The task of identifying generic NPs and coreference resolution are intertwined. In future work, we plan to manually annotate at least part of our corpus with coreference information in order to test to what extent the classification of the pronouns' genericity status can profit from including antecedent information.

8.4.4 Comparison of unigram and bigram models

In this section, we compare the output of the CRF-unigram model to the output of the CRF-bigram model. Specifically, we analyse the cases that the CRF-unigram model gets wrong and the CRF-bigram model gets right, and vice versa. Table 8.23 shows the relevant statistics.

Task NP. As shown in Table 8.23, 9.2% of all instances are labeled wrongly by the unigram model but correctly by the bigram model, and 6.5% are labeled wrongly by the bigram model but correctly by the unigram model. This results in 2.7% improvement of the bigram model over the unigram model. Some of this improvement is "random" in the sense that both models get about 4% generic instances right that the other model gets wrong; but the CRF-bigram model gets significantly more non-generic instances right than the other model. Recall that the WikiGenerics corpus is approximately balanced between generic and non-generic instances, and this effect might turn out differently in a different setting.

	CRF-bigram correct		CRF-unigram correct	
Task NP				
total	948	9.2%	670	6.5%
generic	436	4.2%	407	3.9%
non-generic	512	4.9%	263	2.2%
Task Clause				
total	910	8.7%	575	5.6%
generic	416	4.0%	336	3.2%
non-generic	494	4.8%	236	2.3%
Task Clause+NP				
total	967	9.4%	624	6.0%
GEN_gen	404	3.9%	346	3.3%
NON-GEN_non-gen	532	5.1%	234	2.3%
NON-GEN_gen	31	0.3%	44	0.4%

Table 8.23: Comparison of CRF-bigram to CRF-unigram model: table lists **only** those cases that the respective other model got **wrong**, percentage of all 10355 instances.

Task Cl. A similar effect shows up here as described for Task NP above, though here the bigram model also clearly gets more generic cases right than the unigram model.

Task Cl+NP. The most remarkable difference again, here, is that the bigram model gets more of the non-generic cases right. Regarding the difficult class NON-GEN_gen, the unigram model works better here, as reflected by the F-Scores in Table 8.22.

Qualitative analysis. We now perform a qualitative comparison between the unigram and the bigram model, using the cases that only one of the models got right. Looking at the surrounding clauses can be a "proxy" for coreference resolution as in (7) or (8), which the bigram model labeled correctly, but the unigram model did not.

(7) (a) The invention of the modern piano is credited to Bartolomeo Cristofori (***non-generic***)
(b) who was employed by Ferdinando de' Medici, Grand Prince of Tuscany, as the Keeper of the Instruments; (***non-generic***)
(c) he was an expert harpsichord maker. (***non-generic***)

(8) (a) Pintupi refers to an Australian Aboriginal group (***non-generic***)
(b) who are part of the Western Desert cultural group. (***non-generic***)

There are, however, also cases where a human annotator would have trouble classifying a sentence out of context. In the following, we give several examples for such cases. In all of them, the bigram model outperformed the unigram model. Without further context, clause (9c) has a preference for a generic reading, but one could not unambiguously decide on this annotation; it could also mean "some pacus" inhibit these areas. In the context of the previous clauses, which are clearly generic, however, the reading becomes clear.

(9) (a) A species popular among aquaculturists is the Piaractus mesopotamicus, (***generic***)
(b) also known as "Paraná River Pacu". (***generic***)
(c) Pacus inhabit most rivers and streams in the Amazon and Orinoco river basins of
lowland Amazonia. (***generic***)

Similarly, the last sentence of (10) could be interpreted as talking about a particular screw (e.g., found during some excavation), but the context shows that it is in fact ***generic***.

(10) Archimedes' screw consists of a screw (a helical surface surrounding a central cylindrical shaft) inside a hollow pipe.
The screw is turned usually by a windmill or by manual labour.
As the shaft turns, the bottom end scoops up a volume of water.
This water will slide up in the spiral tube, until it finally pours out from the top of the tube and feeds the irrigation systems.
The screw was used mostly for draining water out of mines or other areas of low lying water.

For the underlined NP in (11), one could also imagine a context in which this describes some particular referent.

(11) Grimpoteuthis is a genus of pelagic umbrella octopus (***generic***)
that live in the deep sea. (***generic***)
Prominent ear-like fins protrude from the mantle just above their lateral eyes. (***generic***)

The underlined NP in (12) could refer to either a particular individual or group, or to a class (which it does in context).

(12) During the summer, narwhals mostly eat Arctic cod and Greenland halibut, (***generic***)
with other fish such as polar cod making up the remainder of their diet. (***generic***)
Each year, they migrate from bays into the ocean as summer comes. (***generic***)

The unigram model wrongly predicted ***non-generic*** for clause (b) in (13), while the bigram model correctly predicted ***generic***. This illustrates that it may also be the case that *following* rather than *preceding* clauses can supply information which results in turning a wrong prediction right.

(13) (a) The study indicated (***non-generic***)
(b) that sloths sleep just under 10 hours a day. (***generic***)
(c) Three-toed sloths are mostly diurnal, (***generic***)
(d) while two-toed sloths are nocturnal. (***generic***)

Example (14) shows a very difficult case for classification, as the subject is only the demonstrative pronoun "this," which could refer to something generic as well if taken out of context. Here, however, it refers to the concrete work done by Koch and is thus non-generic.

(14) In his sixth semester, Koch began to conduct research at the Physiological Institute, (***non-generic***)
where he studied succinic acid secretion. (***non-generic***)
This would eventually form the basis of his dissertation. (***non-generic***)

There are many more examples in the data that would have both generic and non-generic readings when shown out of context. It is comparably harder to find good examples where the context is decisive for assigning the non-generic reading as in (14). Our assumption is that all of the cases that the unigram model gets wrong labeling them generic are in fact random guesses, and the bigram model makes up for this by using the signal from the non-generic context.

8.4.5 Summary

We have presented a novel method for labeling sequences of clauses or their subjects with regard to their genericity, showing that genericity should be treated as a discourse-sensitive phenomenon. Our experiments prove that context information improves automatic labeling results, and that our model outperforms previous approaches by a large margin.

The major contributions of this work include the study of genericity both on the NP- and clause-level, and the study of the interaction of these two levels. Our results of Task Cl+NP show that our model indeed captures the three different types of clauses resulting from the combination of NP-level and clause-level genericity.

At this point, we have presented experiments on identifying the lexical aspectual class of verbs as ***stative*** or ***dynamic***, on distinguishing ***habitual***, ***episodic*** and ***static*** clauses, and on recognizing kind-referring NPs and ***generic*** sentences. We have found a variety of syntactic-semantic and lexical features to be useful for all tasks, and achieved good performance by using Random Forest classifiers and conditional random fields. The methods presented thus far form the basis for our methods for labeling the clauses of a text with their situation entity types, which we present in the next section.

8.5 Automatic classification of situation entity types

In the experiments presented in this section, we are concerned with automatically identifying the **type** of a **situation entity**, which we assume to be expressed by the clauses of a text. Specifically, we present a system for automatically labeling clauses using the inventory of situation entity types shown in Figure 8.2 (details have been described in Section 5.2). We present our experiments on situation entity type classification, beginning with a (near) replication of the system by Palmer et al. (2007), and moving on to evaluate our new approach from multiple perspectives. The experiments in this section are, if not otherwise stated, based on the MASC and Wikipedia (Wiki) data described in Part II of this thesis.

Our approach uses features which, in comparison to prior work, increase robustness: Brown clusters and syntactic-semantic features. In addition, our new model implements the first true sequence labeling model for situation entity types, using conditional random fields to find the globally-best sequence of labels for the clauses in a document. Performance increases by around 2% absolute compared to predicting labels for clauses separately; much of this effect stems from the fact that Generic Sentences cluster together in texts (this has already been discussed in detail in Section 8.4).

State: *The colonel owns the farm.*
Event: *John won the race.*
Report: "...", *said Obama.*
Generic Sentence: Generalizations over kinds.
The lion has a bushy tail.
Generalizing Sentence: Generalizations over events (*habituals*).
Mary often fed the cat last year.
Question: *Who wants to come?*
Imperative: *Hand me the pen!*

Figure 8.2: Situation entity types, adapted from Smith (2003).

8.5.1 Method and baseline

Baseline: UT07. For comparison with prior work, we implement a system similar to the one of Palmer et al. (2007) (henceforth UT07). This system, which we have described in detail in Section 3.1.4, relies on the features summarized in Table 8.24. Feature sets W and T correspond to the set of words and set of POS tags occurring in the data set respectively. We use a minimum frequency threshold of 7 occurrences for each feature. Feature set L comprises sets of predicates assumed to correlate with particular situation entity types, and whether or not the clause contains a modal or finite verb. Set G includes all verbs of the clause and their POS-tags. UT07 additionally uses CCG supertags and grammatical function information. The UT07 system approximates a sequence labeling model by adding the predicted labels of previous clauses as *lookback* (LB) features. To parallel their experiments, we train both MaxEnt and CRF models, as explained in Section 7.2.3.

Feature Set	Explanation
W	words
T	POS tags, word/POS tag combinations
L	linguistic cues
G	grammatical cues

Table 8.24: Features used in baseline UT07.

CRF models. When experimenting with the above explained baseline, we found that words as features are impractical when dealing with larger data sets such as our MASC and Wikipedia corpora. The word features simply capture most of the corpus vocabulary and overfit the model to the data set. In our new model, we address these shortcomings, leveraging features which increase robustness, i.e., distributional information in the form of Brown clusters and a larger set of syntactic-semantic features. We train a conditional random field (CRF), which is able to find the globally-best sequence of labels for the clauses in a document. The UT07 system uses the predicted label of the previous clause or clauses as features, but performing predictions for the clauses in an incremental way. Our system, in contrast, represents the first true sequence labeling model for situation entity types. The system identifies the main verb and main referent of each clause (see Section 7.1.2) and makes use of a variety of features, which we group into two sets:

- Set **A** consists of standard NLP features including part-of-speech (POS) tags and Brown clusters. For an explanation of these features, see Section 7.1.3 and Section 7.1.4.

- Set **B** targets situation entity type labeling, focusing on syntactic-semantic properties of the main verb (**mv**, see Section 7.1.5) – including the linguistic indicators for the main verb (Siegel and McKeown, 2000, see Section 7.1.6) – and the main referent (**mr**, see Section 7.1.7), as well as properties of the clause which indicate its aspectual nature (**cl**, see Section 7.1.8).

Comparison with baseline. Palmer et al. (2007) have created a data set of 20 texts from the *popular lore* section of the Brown corpus (Francis and Kučera, 1979), manually segmented into 4391 clauses and marked by two annotators in an intuitive way with κ=0.52 (Cohen, 1968). Final labels were created via adjudication. The texts are essays and personal stories with topics ranging from maritime stories to marriage advice.

Results on the Brown data, with the same training/test split as in the original work, appear in Table 8.25. Unlike their maximum entropy model using lookback features, our CRF predicts the label sequence jointly and outperforms UT07 on the Brown data by up to 7%

	Palmer et al. (2007)		Our implementation	
Features	**MaxEnt**	**LB**	**MaxEnt**	**CRF**
W	45.4	46.6	48.8	47.0
WT	49.9	52.4	52.9	53.7
WTL	48.9	50.5	51.6	55.8
WTLG	**50.6**	**53.1**	**55.8**	**60.0**

Table 8.25: Accuracy on Brown. Test set majority class (State) is 35.3%. LB = results for best lookback settings in MaxEnt. 787 test instances.

accuracy. We assume that the performance boost in the MaxEnt setting is at least partially due to having better parses.

In sum, on the small Brown data set, a CRF approach successfully leverages sequence information, and a simple set of features works well. Preliminary experiments applying our new features on Brown data yield no improvements, suggesting that word-based features overfit this domain.

8.5.2 Impact of feature sets

We now turn to the our data set of MASC and Wikipedia texts labeled with their situation entity type (see Chapter 6). We train a CRF for labeling situation segments with their type, experimenting with two feature sets which each contain several feature groups. Feature set **A** consists of syntactic information in the form the part-of-speech tags occurring in a segment (Section 7.1.3) and the Brown clusters for each word in a segment (Section 7.1.4). Feature set **B**, in contrast, contains features targeted at the situation entity classification task. **B** contains features describing the clause's main referent (**mr**, see Section 7.1.7). Additional **B** features describing the verb constellation involving the clause's main verb (**mv**, see Section 7.1.5), including linguistic indicator features for the main verb's lemma (see Section 7.1.6). Finally, **B** contains features describing aspects of the clause (**cl**, see Section 7.1.7), such as negation or other modifiers.

Table 8.26 shows the results for 10-fold CV on the dev part of the MASC+Wiki corpus (see Section 8.1 for experimental settings). Each feature set on its own outperforms the majority class baseline. Of the individual feature groups, **bc** and **mv** have the highest predictive power; both capture lexical information of the main verb. Using sets **A** and **B** individually results in similar scores; their combination increases accuracy on the dev set by an absolute 3.6-4.3%. Within **A** and **B**, each subgroup contributes to the increase in performance (not shown in table).

Finally, having developed exclusively on the dev set, we run the system on the held-out test set, training on the entire dev set. Results (in Table 8.27) show the same tendencies as for the dev set: each feature set contributes to the final score, and the syntactic-semantic features targeted at classifying situation entity types (i.e. **B**) are helpful.

Feature set	**Precision**	**Recall**	**F1**	**Accuracy**
majority class (State)	6.4	14.3	8.8	45.0
A	70.1	61.4	65.4	∗†72.1
pos	49.3	40.3	44.3	58.7
bc	67.5	55.8	61.1	∗70.6
B	69.5	62.7	66.9	⋆‡72.8
mr	36.4	26.8	30.9	51.7
mv	62.3	52.4	56.9	⋆70.8
cl	53.3	41.2	46.6	52.8
A+B	**74.1**	**68.6**	**71.2**	‡†**76.4**
upper bound (humans)	78.6	78.6	78.6	79.6

Table 8.26: Impact of different feature sets. Wiki+MASC dev set, CRF, 10-fold CV.

Feature set	**Precision**	**Recall**	**F1**	**Accuracy**
majority class (State)	6.4	14.3	8.8	44.7
A	67.6	60.6	63.9	∗69.8
B	69.9	61.7	65.5	†71.4
A+B	**73.4**	**65.5**	**69.3**	∗†**74.7**

Table 8.27: Results on MASC+Wiki **held-out test set** (7937 test instances).

Feature set	Precision	Recall	F1	Accuracy
A+B	**74.1**	**68.6**	**71.2**	**76.4**
- bc	71.3	65.7	68.4	74.5
- pos	73.4	67.4	70.2	76.0
- mr	73.7	67.4	70.4	75.9
- mv	72.3	64.2	68.0	73.6
- cl	73.1	67.6	70.2	76.0

Table 8.28: Impact of feature groups: ablation Wiki+MASC dev set, CRF, 10-fold CV. All accuracies for ablation settings are significantly different from A+B.

Ablation. To gain further insight, we ablate each feature subgroup from the full system, see Table 8.28. Again, **bc** features and **mv** features are identified as the most important ones. The other feature groups partially carry redundant information when combining **A** and **B**. Next, we rank features by their information gain with respect to the situation entity types. In Table 7.1, the features of each group are ordered by this analysis. Ablating single features from the full system does not result in significant performance losses. However, using only selected, top features for our system decreased performance, possibly because some features cover rare but important cases, and because the feature selection algorithm does not take into account the information features may provide regarding transitions (Klinger and Friedrich, 2009). In addition, CRFs are known to be able to deal with a large number of potentially dependent features.

Side remark: pipeline approach. Feature set **B** is inspired by previous work on two subtasks of assigning an situation entity type to a clause: (a) identifying the genericity of a noun phrase in its clausal context (see Section 8.4), and (b) identifying whether a clause is episodic, habitual or static (see Section 8.3). This information can in turn be used to determine a clause's situation entity type label in a rule-based way, e.g., Generalizing Sentences are habitual clauses with a non-generic main referent. As our corpus is also annotated with this information, we also trained separate models for these subtasks and assigned the situation entity type label accordingly. However, such a pipeline approach is not competitive with the model trained directly on situation entity types. We here use the subset of situation entities labeled as Event, State, Generic Sentence or Generalizing Sentence because noun phrase genericity and habituality is not labeled for Imperative and Question, and Report is identified lexically based on the main verb rather than these semantic features. Models for subtasks of situation entity type identification, i.e., (a) genericity of noun phrases and (b) habituality reach accuracies of (a) 86.8% and (b) 83.6% (on the relevant subset). Applying the labels output by these two systems as (the only) features in a rule-based way using a J48 decision tree (Hall et al., 2009) results in an accuracy of 75.5%, which is lower than 77.2%, the accuracy of the CRF which directly models situation entity types (when using only the above four types).

8.5.3 Impact of amount of training data

Next we test how much training data is required to get stable results for situation entity type classification. Figure 8.3 shows accuracy and F1 for 10-fold CV using **A+B**, with training data downsampled to different extents in each run by randomly removing documents. Up to the setting which uses about 60% of the training data, performance increases steadily. Afterwards, the curves start to level off. We conclude that robust models can be learned from our corpus. Adding training data, especially in-domain data, will, as always, be beneficial.

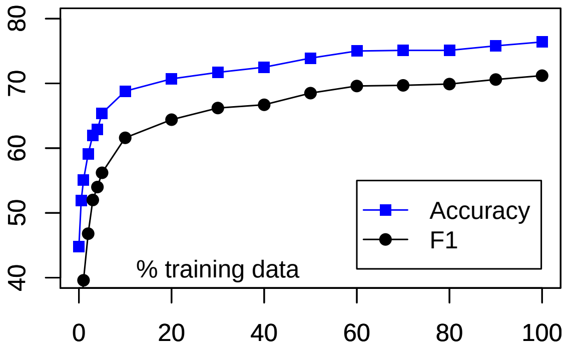

Figure 8.3: Learning curve for MASC+Wiki dev.

8.5.4 Impact of sequence labeling approach

Palmer et al. (2007) suggest that situation entity types of nearby clauses are a useful source of information. We further test this hypothesis by comparing our sequence labeling model (CRF) to two additional models: (1) a MaxEnt model, which labels clauses in isolation, and (2) a MaxEnt model including the correct label of the preceding clause (seq-oracle), simulating an upper bound for the impact of sequence information on our system.

Table 8.29 shows the results. Scores for Generalizing Sentence are the lowest as this class is very infrequent in the data set. The most striking improvement of the two sequence labeling settings over MaxEnt concerns the identification of Generic Sentences. These often "cluster" in texts (Friedrich and Pinkal, 2015a) and hence their identification profits from using sequence information. The results for seq-oracle show that the sequence information is useful for State, Generic and Generalizing Sentences, but that no further improvement is to be expected from this method for the other situation entity types. We conclude that the CRF model is to be preferred over the MaxEnt model; in almost all of our experiments it performs significantly better or equally well.

8.5.5 Impact of genre

In this section, we test to what extent our models are robust across genres. Table 8.30 shows F1-scores for each situation entity type for two settings: the 10-fold CV setting as explained in section 8.1, and a **genre-CV** setting, simulating the case where no in-genre

Situation entity type	MaxEnt	CRF	seq-oracle
State	79.1	**80.6**	81.7
Event	77.5	**78.6**	78.3
Report	78.2	**78.9**	78.3
Generic	61.3	**68.3**	73.5
Generalizing	25.0	**29.4**	38.1
Imperative	72.3	**75.3**	74.7
Question	84.4	**84.4**	83.8
macro-avg P	71.5	**74.1**	75.5
macro-avg R	66.1	**68.6**	70.4
macro-avg F1	68.7	**71.2**	73.9
accuracy	∗74.1	∗†**76.4**	†77.9

Table 8.29: Impact of sequence information: (**F1** by situation entity type): CRF, Masc+Wiki, 10-fold CV.

training data is available, treating each genre as one cross validation fold. As expected, in the latter setting, both overall accuracy and macro-average F1 are lower compared to the case when in-genre training data is available. Nevertheless, our model is able to capture the nature of situation entity types across genres: the prediction of State, Event, Report and Question is relatively stable even in the case of not having in-genre training data. An Event seems to be easily identifiable regardless of the genre. Generic Sentence is a problematic case; in the genre-CV setting, its F1-score drops by 23.5%. The main reason for this is that the distribution of situation entity types in Wikipedia differs completely from the other genres (see section Section 6.7). Precision for Generic Sentence is at 70.5% in the genre-CV setting, but recall is only 32.8% (compared to 70.1% and 66.6% in the 10-fold CV setting). Genericity seems to work differently in the various genres: most generics in Wikipedia clearly refer to kinds (e.g., lions or plants), while many generics in essays or letters are instances of more abstract concepts or generic *you*.

Results by genre. Next, we drill down in the evaluation of our system by separately inspecting results for individual genres. Table 8.31 shows that performance differs greatly depending on the genre. In some genres, the nature of situation entity types seems clearer to our annotators than in others, and this is reflected in the system's performance. The majority class is Generic Sentence in wiki, and State in all other genres. In the 'same genre' setting, 10-fold CV was performed within each genre. Adding out-of-genre training data improves macro-average F1 especially for genres with low scores in the 'same genre' setting. This boost is due to adding training data for types that are infrequent in that genre. Accuracy (not shown in table) improves significantly for blog, essays, govt-docs, jokes, and journal, and does not change for the remaining genres. We conclude that it is extremely beneficial to use our full corpus for training, as robustness of the system is increased, especially for situation entity types occurring infrequently in some genres.

Situation entity type	genre-CV	10-fold CV	Humans
State	78.2	80.6	82.8
Event	77.0	78.6	80.5
Report	76.8	78.9	81.5
Generic	44.8	68.3	75.1
Generalizing	27.4	29.4	45.8
Imperative	70.8	75.3	93.6
Question	81.8	84.4	90.7
macro-avg F1	66.6	71.2	78.6
accuracy	*71.8	*76.4	79.6

Table 8.30: Impact of in-genre training data. F1-score by situation entity type, CRF, MASC+Wiki dev.

8.5.6 Identification of Abstract Entities

Our system as described in above notably does not address one of Smith's main situation entity categories: Abstract Entities, introduced in detail in Section 5.2.3. These situation entities are expressed as clausal arguments of certain predicates such as (canonically) "know" or "believe" as illustrated by the examples in Figure 8.4. For a more detailed description of Abstract Entities, see Section 1.1 and Section 5.2.3.

Fact: Objects of knowledge.	*I know that Mary refused the offer.*
Proposition: Objects of belief.	*I believe that Mary refused the offer.*

Figure 8.4: Abstract Entity situation entity types.

During the creation of the corpus, annotators were asked to give one label out of the SE types included in our classification task, and to mark the clause with one of the Abstract Entity subtypes in addition if applicable. Analysis of the data shows that our annotators frequently forgot to mark clauses as Abstract Entities, which makes it difficult to model these categories correctly. As a first step toward resolving this issue, we implement a filter which automatically identifies Abstract Entities by looking for clausal complements of certain predicates. The list of predicates is compiled using WordNet synonyms of *know*, *think*, and *believe*, as well as predicates extracted from FactBank (Sauri and Pustejovsky, 2009) and TruthTeller (Lotan et al., 2013). Many of the clauses automatically identified as Abstract Entities are cases that annotators missed during annotation. We thus performed a post-hoc evaluation, presenting these clauses in context to annotators and asking whether the clause is an Abstract Entity. The so-estimated precision of our filter is 85.8% (averaged over 3 annotators). Agreement for this annotation task is $\kappa = 0.54$, with an observed agreement of 88.7%. Our filter finds 80% of the clauses labeled as Abstract Entity by at least one annotator in the gold standard; this is *approximately* its recall.

	Training data				
	majority	same			
	class	genre	all	Humans	
Genre	**%**	**F1**	**F1**	**F1**	κ
blog	57.6	57.3	**64.9**	72.9	0.62
email	68.6	63.6	**66.4**	67.0	0.65
essays	49.4	33.5	**62.1**	64.6	0.54
ficlets	44.7	60.2	**65.7**	81.7	0.80
fiction	45.8	63.0	**66.0**	76.7	0.77
govt-docs	60.9	26.6	**67.6**	72.6	0.57
jokes	34.9	66.2	**69.8**	82.0	0.77
journal	59.3	35.8	**59.8**	63.7	0.52
letters	57.3	51.9	**65.1**	68.0	0.66
news	52.2	54.6	**64.1**	78.6	0.75
technical	57.7	31.4	**59.4**	54.7	0.55
travel	25.9	39.9	**58.1**	48.9	0.59
wiki	51.6	53.1	**63.0**	69.2	0.66

Table 8.31: Macro-avg. F1 by genre, CRF, 10-fold CV. Majority class given in % of clauses.

8.5.7 Summary

We have presented a system for automatically labeling clauses with their situation entity type which is mostly robust to changes in genre and which reaches accuracies of up to 76%, comparing favorably to the human upper bound of 80%. The system benefits from capturing contexual effects by using a linear chain CRF with label bigram features. In addition, the distributional and targeted syntactic-semantic features we introduce enable situation entity type prediction for large and diverse data sets. The feature sets include information that has been shown to be useful for sub-tasks of situation entity type classification, specifically for identifying the lexical aspectual class of a clause's main verb, the genericity of its main referent and habituality of the clause.

8.6 Conclusion

This chapter has provided an experimental evaluation of our computational models for labeling the clauses of a text with situation entity types and related aspectual and semantic information. The labeled corpus that is the basis for our experiments is much larger than any of the data sets used in previous related research, and covers a variety of genres and domains. This allows for the following conclusions.

We have confirmed that a semi-supervised approach following ideas of Siegel and McKeown (2000) to classifying verbs as ***stative*** or ***dynamic*** works well. Our experiments show that the corpus-based linguistic indicator patterns generalize from one verb type to another, which means that they can be used to predict the aspectual class of verbs not seen in the labeled training set. In addition, we have shown that features capturing the local syntactic and semantic context of a verb are necessary in order to improve upon the majority baseline which simply assigns the most frequent class to each verb type that has been seen in the training set. For verbs that occur with only one of the labels in the training set, an accuracy of 93% is reached by the majority class baseline; accuracy for verbs that are ambiguous, i.e., which occur both as ***stative*** and as ***dynamic*** (or as ***both***) in our corpus, is only around 80%, but our system outperforms the majority class baseline in this case.

In the next step, we have created the first fully automatic approach for classifying *all* clauses of a text (rather than just a manually selected sample of sentences, as in work by Mathew and Katz (2009)) with respect to their aspectual properties as ***habitual***, ***episodic*** or ***static***. We have found syntactic-semantic features capturing properties of the clause to be essential in addition to type-based linguistic indicator features. The best results for the three-way classification task were achieved using a cascaded approach, which first identifies ***static*** clauses and then classifies the remaining clauses into ***habitual*** and ***episodic***. Compared to a jointly trained model using three classes at once, this cascaded approach increases precision and accuracy especially for the least frequent class ***habitual***. On our Wikipedia data, an accuracy of 80% is reached, compared to a majority class baseline of 60%.

Our third set of experiments addressed the recognition of genericity with a focus on identifying subjects that refer to a kind. On the ACE corpora and on our Wikipedia data set, we outperform previous work (Reiter and Frank, 2010) by using a discourse-sensitive sequence labeling method, which correctly classifies many difficult cases where the information in the clause itself is not sufficient to assign a label. In addition, our models are the first to address the distinction between clauses making a generic statement ("Blobfish are ugly") and clauses reporting on a particular situation involving the reference to a kind ("In September 2013, the blobfish was voted 'The World's Ugliest Animal.'"). Our best model reaches an accuracy of 77.4%, compared to a discourse-unaware method, which achieves 74.0%, and a majority class baseline resulting in 50.4% accuracy.

Finally, we present the first true sequence labeling model for situation entity types. We build on the features and methods identified as effective by the experiments described above, each of which addresses a sub-task of the situation entity classification task. Our system, inspired by previous work by Palmer et al. (2007), incorporates the syntactic-semantic and type-based features which we found useful for capturing aspectual distinctions and genericity. In contrast to the previous work, our model performs robustly across genres and domains by using distributional information in the form of Brown clusters as features, instead of words. Our best model achieves an accuracy of 76.4%, which compares favorably to a majority class baseline of 45.0% and an upper bound (as determined by human annotator agreement) of 79.6%.

Part IV

Further directions and conclusion

Chapter 9

Situation entity types and temporal relations

In Chapter 1, we have motivated our work on identifying the aspectual nature of clauses by the different contributions that eventuality types make to the temporal interpretation of discourse. When two bounded events occur sequentially in discourse, in absence of further markers, the default interpretation is that the first one occurs after the second (Smith and Erbaugh, 2005). For example, in (1), (b) is interpreted as happening right after (a).

(1) (a) Sue entered the restaurant. (Event)
(b) She joined her friends at the bar. (Event)

In contrast, if one of the sentences is a State as in (2), the default interpretation is that the two situations happen simultaneously or at least overlap.

(2) (a) Sue entered the restaurant. (Event)
(b) Her friends were sitting at the bar. (State)

In this section, we present two corpus studies indicating that these patterns can be found in corpora. We automatically label the texts of the Penn Discourse TreeBank (Prasad et al., 2008) and TimeBank (Pustejovsky et al., 2006) with situation entity types, using our best model as described in Section 8.5, trained on the entire development part of our MASC and Wikipedia corpus. This classifier has an accuracy of around 76%, which is high enough to be able to show tendencies. In the first corpus study, we find that PDTB Temporal relations exhibit different patterns of situation entity types in their arguments compared to other PDTB discourse relations, and that the three PDTB Temporal relations also differ with respect to the distribution of situation entity types that occur in their argument spans. In the second corpus study, we show that the differences in situation entity type distribution are also visible for different temporal relations in TimeBank. It follows that situation entity types capture information about the differences between the temporal relations. In contrast, this is not true for the semantic event class labels that are part of the TimeML annotations.

Heuristic identification of main verbs. For practical reasons, we here apply a heuristic situation segmentation method that, in contrast to the method described in Chapter 4, does not rely on SPADE (Soricut and Marcu, 2003). Our heuristic segmenter makes use of the part-of-speech tags and parse trees generated by the Stanford parser (Klein and Manning, 2002). We label each verb with its tense, grammatical aspect and voice using the method of Loaiciga et al. (2014). We identify the following verbs as main verbs of situation segments: verbs labeled as having a finite tense and gerunds that function as introducing reduced relative clauses.[1] We include all tokens that are grammatical dependents of the main verb into the heuristically identified "segment" and compute all feature values based on this set of tokens.

9.1 Comparison with discourse relations in PDTB

The Penn Discourse TreeBank (PDTB, Prasad et al., 2007, 2008) provides a theory-neutral lexically-grounded annotation of both Explicit and Implicit discourse relations over the Wall Street Journal section of the Penn TreeBank (Marcus et al., 1993). Explicit relations are annotated whenever a *discourse connective* such as "because" or "while" is present in the text. The annotator's task is to chose the two spans of the texts that are semantically connected by the connective, i.e., the *arguments* of the discourse relation. For a discussion on the extents that these spans take, see Section 4.2. The two spans are called ARG1 and ARG2. In Explicit relations, ARG2 is the argument to which the connective is syntactically bound, as illustrated by examples (3) and (4), which are taken from Prasad et al. (2008).

(3) [The Mountain View, Calif., company has been receiving 1,000 calls a day about the product $_{ARG1}$] since [it was demonstrated at a computer publishing conference several weeks ago. $_{ARG2}$] (**Temporal**)

(4) [It was a far safer deal for lenders $_{ARG1}$] since [NWA had a healthier cash flow and more collateral on hand. $_{ARG2}$] (**Causal**)

Each relation is assigned one or two relation senses from a hierarchy with three levels. The four top-level senses are Temporal, Contingency, Comparison and Expansion. The full set of relations contains many relation senses that occur very infrequently; we thus here work with the subset of PDTB relations that was used in the CoNLL shared task on shallow discourse parsing (Xue et al., 2015). The set of relations is listed in Figure 9.1.

In addition to Explicit relations, PDTB annotates Implicit, EntRel and AltLex relations. Implicit relations are marked between adjacent sentences if no discourse connective is present. The first sentence is ARG1 and the second sentence is ARG2 in this case. Annotators are asked to "insert" a lexical discourse connective and then give a relation sense

[1] Technically, we identify the heads of reduced relative clauses in the following way. We mark verbs having the part-of-speech tag `VBG` and that are the head of `vmod`, `xcomp` or `ccomp` dependency relation, as well as verbs tagged as `VBN` that are not head of a `amod` relation.

in the same way as for Explicit relations. EntRel is used if only an entity-based coherence relation can be perceived between the two sentences. AltLex marks cases where the discourse relation is lexicalized, though by means other than discourse connectives. AltLex relations are for instance signaled by phrases such as "the reason was," and they are assigned one of the relation senses.

Situation entity types of PDTB relation arguments. In the first part of this corpus study, we investigate whether the different PDTB relations show different patterns of situation entity types in their arguments. We use all cases in this study where the main verb of at least one of the PDTB arguments was labeled with a situation entity type by our automatic classifier. Figure 9.1 shows the resulting distributions of situation entity types per PDTB relation sense. In all senses except the Temporal relations, the distribution of situation entity types looks similar and seems to follow the overall distribution of situation entity types that can be expected over a large volume of text: STATE is most frequent, followed by EVENT, and there are only few GENERIC SENTENCES and GENERALIZING SENTENCES.

For Temporal relations, on the other hand, both Explicit and Implicit relations show differences in the distributions of situation entity types. In general, Temporal relations have a much higher frequency of EVENTS, with the exception of STATE being the most frequent type for Succession in the case of Implicit relations.

We also compared the distributions of situation entity types of ARG1 and ARG2 for each relation. For most relations, we do not observe a different tendency in the distribution of situation entity types between ARG1 and ARG2 (not shown in the table). Notable exceptions are only the Temporal relations and the Contingency.Condition relation. In the case of Implicit relations, all ARG2s of Contingency.Condition relations are labeled as STATE, which is not surprising due to our annotation guidelines (conditionals are all labeled as STATES). The ARG1s of Contingency.Condition relations consist of 50% STATES and 50% GENERIC SENTENCES. To sum up, we find that there are patterns based on situation entity types that help to distinguish Temporal relations from other PDTB discourse relations. We next drill down on the patterns that can be observed in the case of Temporal relations.

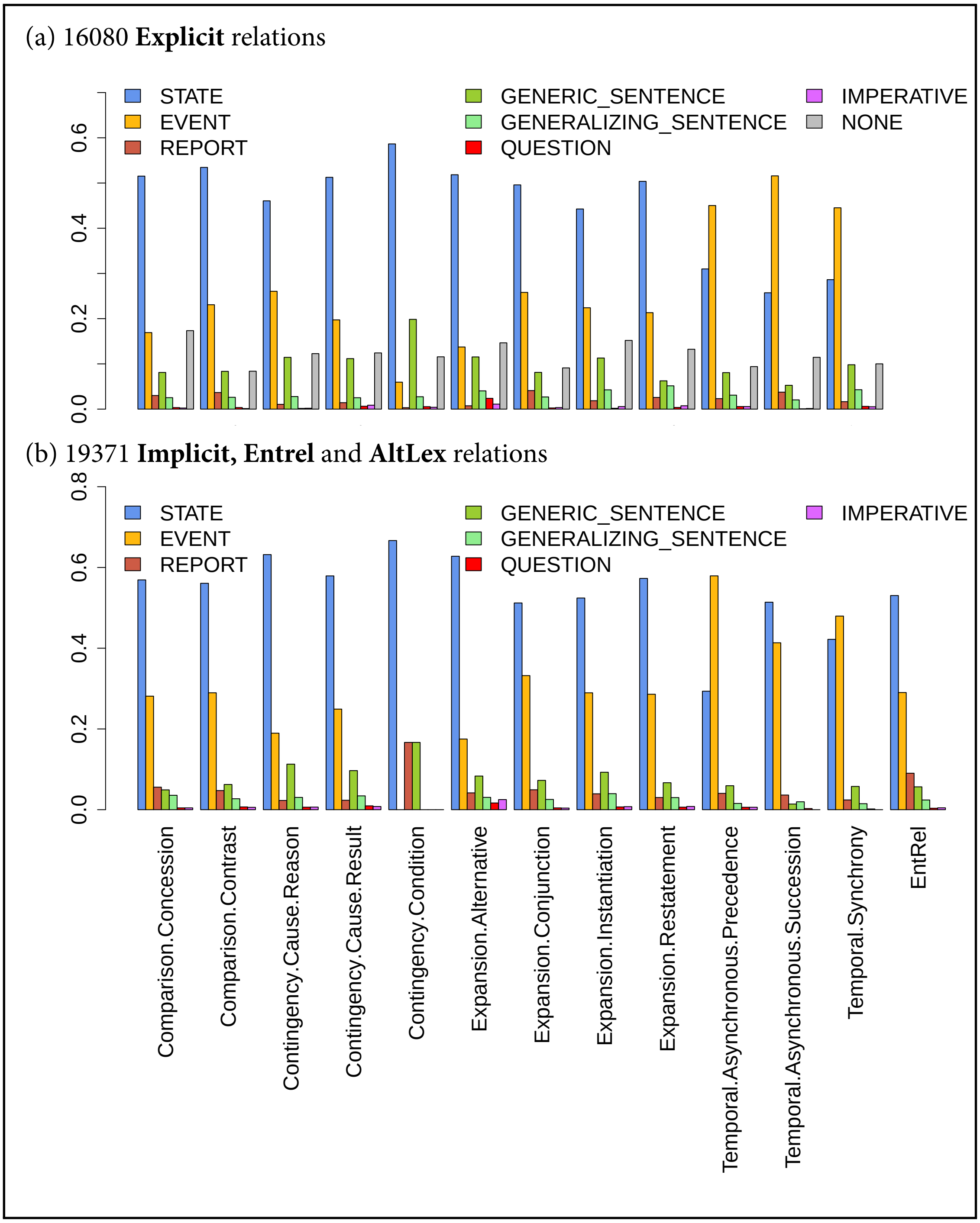

Figure 9.1: Predicted situation entity types of PDTB arguments (ARG1 & ARG2), distributions normalized per PDTB relation type: entire PDTB, Explicit vs. Implicit relations.

Situation entity types and PDTB Temporal relations. Figure 9.2 shows the distributions of the situation entity types for ARG1 and ARG2 of the three Temporal relations in PDTB. We show these distributions for Explicit and Implicit relations separately. The relation type Synchrony is used when the situations described in ARG1 and ARG2 overlap, as in (5). Precedence is used for cases where the situation described in ARG1 is temporally located before the situation in ARG2 (6); for Succession ARG2 happens before ARG1 (7).

(5) [While many of the risks were anticipated $_{ARG1}$ Event] when [Minneapolis-based Cray Research first announced the spinoff in May. $_{ARG2}$ Event] (Explicit, **Temporal.Synchrony**)

(6) [The man, whom it did not name, had been found to have the disease after hospital tests. $_{ARG1}$] (Implicit: then) [Once the disease was confirmed, all the man's associates and family were tested, but none have so far been found to have AIDS. $_{ARG2}$] (**Temporal.Asynchronous.Precedence**)

(7) When [the company asked members in a mailing which cars they would like to get information about for possible future purchases $_{ARG2}$ Event] [Buick came in fourth among U.S. cars and in the top 10 of all cars $_{ARG1}$ Event] (Explicit, **Temporal.Asynchronous.Succession**)

We observe in Figure 9.2 that the ratio of States and Events differs across relations and that it also differs for Implicit vs. Explicit relations. In Explicit relations, the most frequent situation entity type in ARG2 is Event for all three Temporal relations. A possible explanation is that discourse connectives are used to explicitly state the temporal relationship of an Event in ARG2 to any other type of situation. Implicit relations, in contrast, rely on aspectual information to a greater extent. The distributions of situation entity types of ARG1 and ARG2 look different from each other within and across relations: In the case of Precedence, Event is the most frequent type in both ARG1 and ARG2; in Succession, State is more frequent, especially in ARG2. Synchrony relations use State more often as the situation entity type of one of their arguments, which will lead to an interpretation of situations being temporally overlapping or parallel (see the introductory example (2) above). Precedence uses predominantly Event for both arguments – the default interpretation of two events mentioned consecutively in discourse is that they happen in this order (see the discussion of (1) above). The interpretation that the second sentence (ARG2) happens before the first one (ARG1) is more likely if at least one of the arguments is a State, as illustrated by example (8).

(8) [Mr. Steinberg is thought to be on friendly terms with UAL's Mr. Wolf. $_{ARG1}$ State] (Implicit, earlier) [The investor was instrumental in tapping Mr. Wolf to run the air cargo unit of Tiger International Inc. $_{ARG2}$ State] (**Temporal.Asynchronous.Succession**)

In sum, Temporal relations in PDTB exhibit clear preferences for certain types of situation entities in their respective ARG1 and ARG2. This is not the case for other PDTB discourse relations. In this study, we do not drill down further on the patterns observed for Temporal relations for reasons of data sparsity. However, the observations presented here point into directions for promising future research.

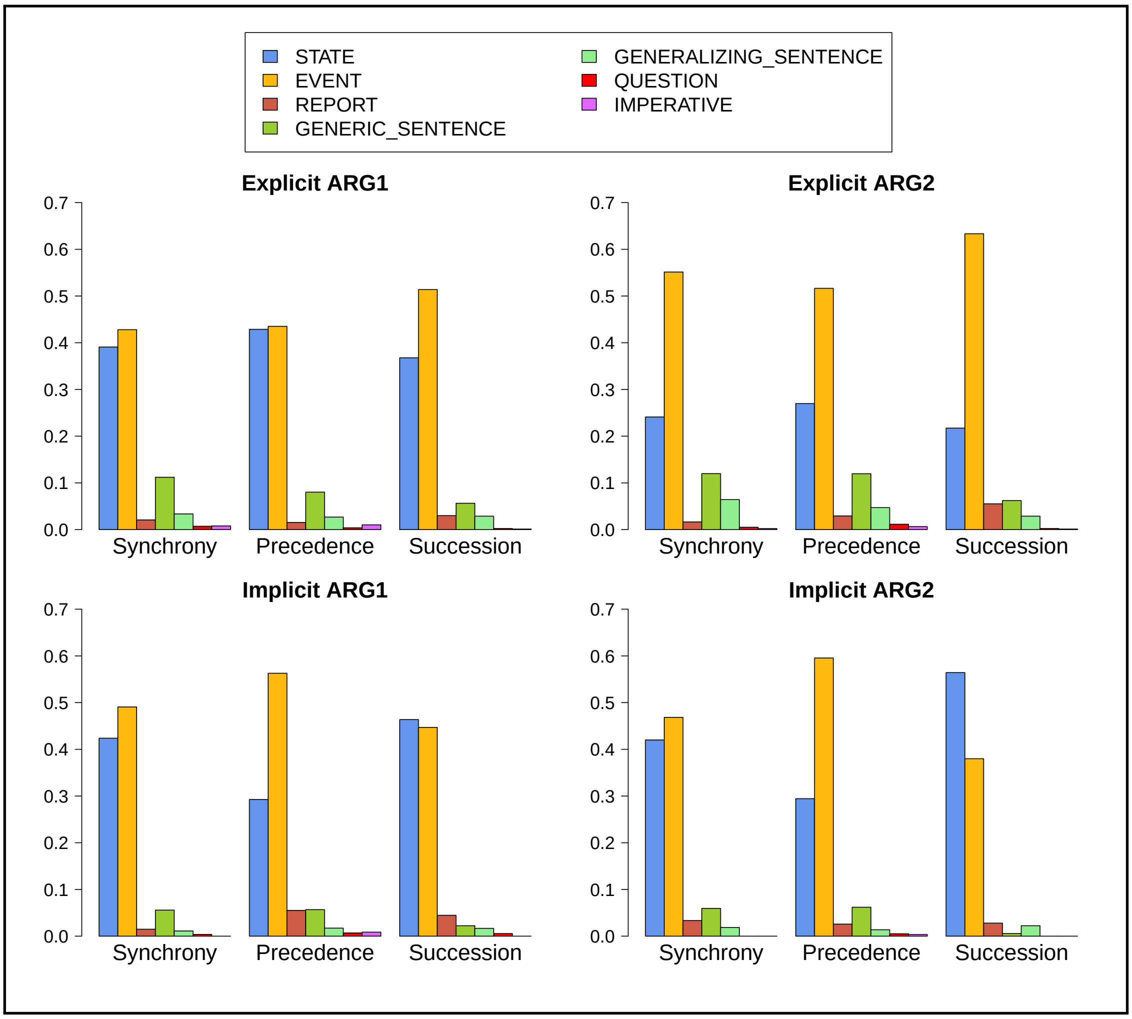

Figure 9.2: Predicted situation entity types of PDTB arguments for Temporal relations, normalized per relation, entire PDTB.

9.2 Comparison with TimeBank relations

In this section, we present a corpus study comparing the annotations given in TimeBank 1.2 (Pustejovsky et al., 2006) to automatically predicted situation entity type labels.[2] TimeBank 1.2 contains 183 articles from the news domains that are labeled according to the TimeML annotation scheme (Pustejovsky et al., 2003a). This includes the tagging of verbs and other constructions as events, the identification of temporal expressions as TIMEX and the labeling of links between events and TIMEXes or between events and

[2] TimeBank contains several documents from the Wall Street Journal (WSJ) section of the Penn TreeBank which are also part of the MASC development data used to train our situation entity classifier. We did not remove those documents from the training data as the goal of this corpus study was not to estimate performance of the classifier, but to compare situation entity type labels to the annotations in TimeBank.

events as temporal relations (TLINKs). Each event is assigned a "semantic event class," which roughly corresponds to situation types. For example, OCCURRENCE captures "events that happen," STATE is used for describing "circumstances in which something holds," and PERCEPTION is used for events involving the physical perception of another event, including verbs such as "see" or "hear." For a comparison of TimeML event classes and situation entity types, see Section 3.1.

In this corpus study, we predict situation entity types for verbs tagged as events in TimeBank and compare them to the manual event class labels given in TimeBank. Second, for each TimeML temporal relation between two events (called event 1 and event 2 as they occur in linear order in the text), we compare the patterns of situation entity types and the patterns of TimeML event classes. We show (a) that the two label sets do not correspond to each other and (b) that situation entity types are more appropriate than TimeML event classes for capturing differences between the different temporal relations.

We conduct this corpus study on the subset of 2525 temporal relations (TLINKs) between events that were assigned a situation entity type label by our automatic classifier. We use all event annotations that could be mapped to the main verb of a situation segment. (TimeBank marks a larger set of syntactic constructions as events, e.g., nouns or infinitives.)

First, we plot the distributions of situation entity types per TimeML event class (Figure 9.3). We can see that the TimeML event types do not correspond to the situation entity types; while we see more States for STATE and I_STATE and comparably more Events for OCCURRENCE and I_ACTION, State and Event occur to a substantial extent for all TimeML event classes. As expected, REPORTING and Report cover the same cases.

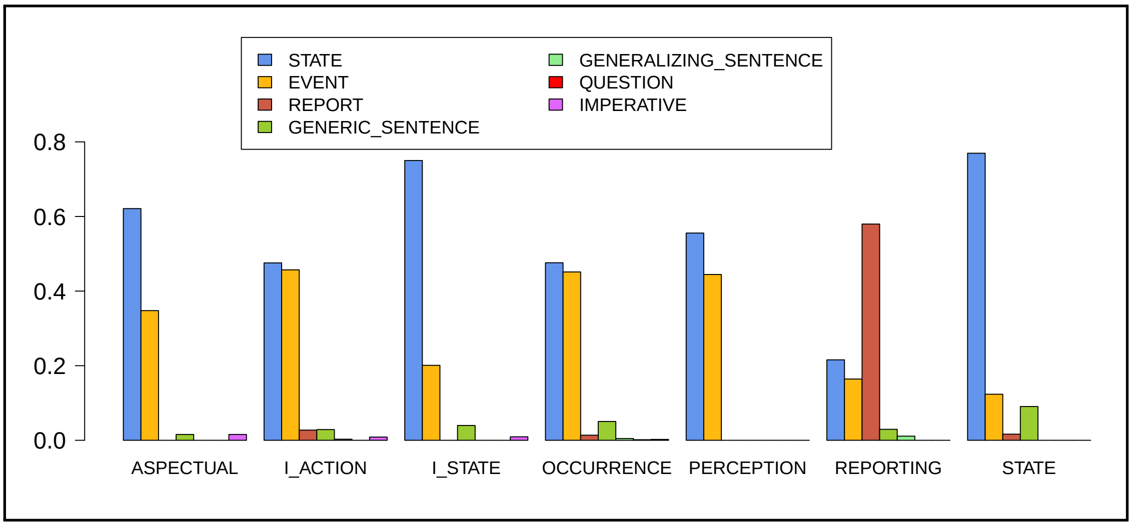

Figure 9.3: TimeML EVENT types versus situation entity types.

Next, we address the question of how well TimeML event classes capture the difference between the TimeML temporal relations. We use the relations as annotated in TimeBank-Dense (Cassidy et al., 2014). In this corpus, temporal relations are annotated for all pairs of events in a document. While the original TimeBank uses a set of 14 relations following Allen (1981), TimeBank-Dense uses only the six relations *before*, *after*, *includes*, *is_included*, *simultaneous* and *vague*. The latter type of relation is assigned whenever a majority vote could not be reached by the two or three annotators that label each document.

Apparently, TimeML event types are not a strong indicator of the different TimeML relations. Figure 9.4 shows that OCCURRENCE is the most frequent event class in all cases except for the first event in *is-included* relations. The distributions for events 1 and 2 look more or less the same for both *after* and *before* relations, the most frequent event class being OCCURRENCE in each case. STATEs are more frequent as the second event of *is_included* relations, yet, OCCURRENCE is still more frequent. REPORTING is the only event class that seems to differ across relations: it occurs more frequently as the first event in *is_included* relations or, correspondingly, as the second event in *includes* relations. However, this is an artifact of how attributions are annotated in the text rather than a temporal pattern.

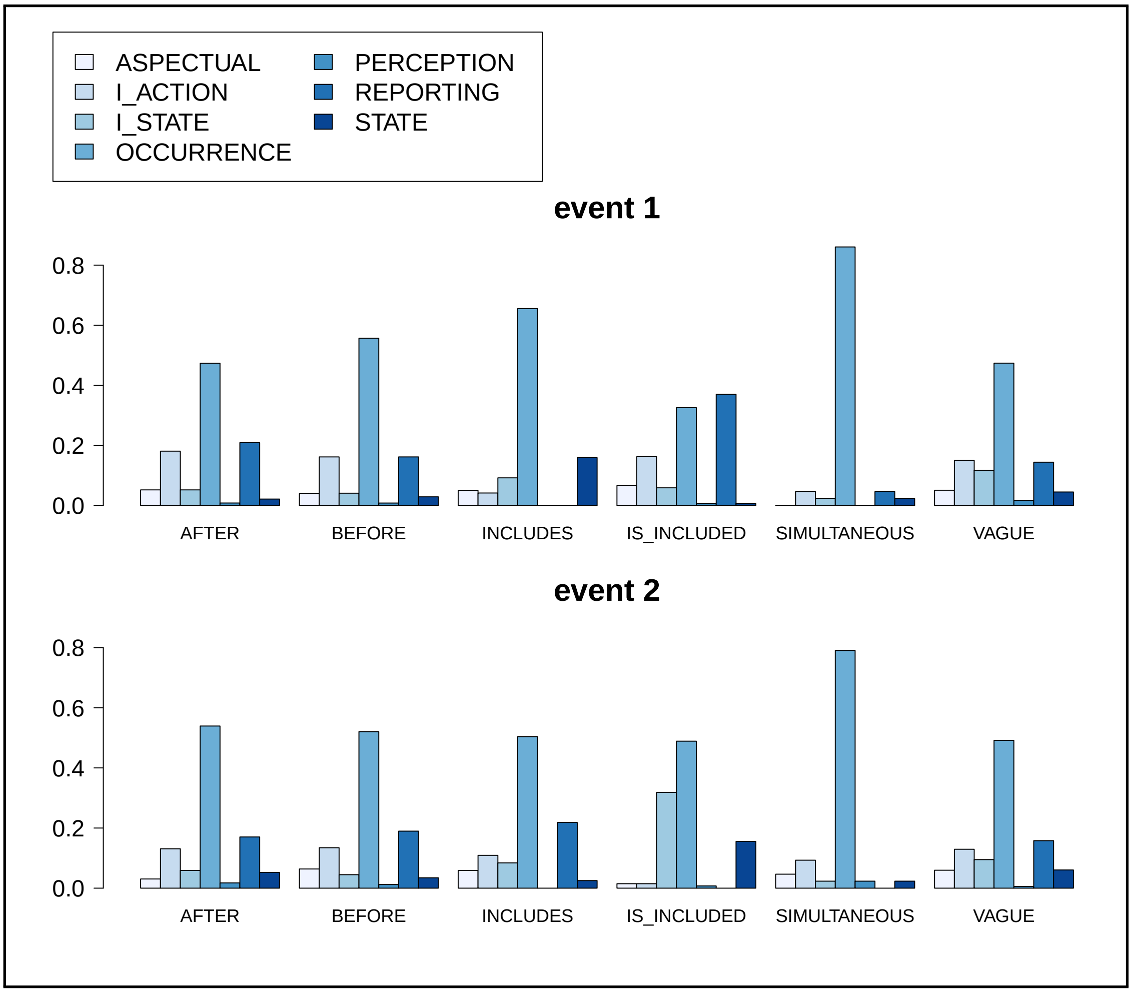

Figure 9.4: TimeML event types versus TimeML temporal relations (event-event relations).

We conduct a parallel study for the same set of TimeML event-event relations, plotting the distributions of situation entity types per relation. As can be seen in Figure 9.5, situation entity types show differences between temporal relations more clearly than the original TimeML event classes. *After* and *before* relations clearly differ from each other, and event

1 and event 2 show similar patterns which correspond to each other. Specifically, event 1 of *after* and event 2 of *before* show a larger percentage of State, while event 1 of *before* and event 2 of *after* have more Events. Similarly, event 1 and 2 are inversed for the *includes* and *is_included* relations, with States being more frequent for the event that includes the other one, event 1 in the former and event 2 in the latter case.

Events of *simultaneous* relations are marked predominantly for OCCURRENCE in TimeBank and as Event by our situation entity classifier. However, a larger percentage of situations are marked as State. Finally, the distribution of TimeML event classes between *vague* relations did not show any pattern different from those corresponding to other temporal relation types. In the case of situation entity types, they look clearly distinct from the other relations, with the most frequent type of both event 1 and event 2 being a State. A possible explanation is that many of these States are coerced cases, e.g., negated or modalized cases, which were thus hard to classify for temporal relations.

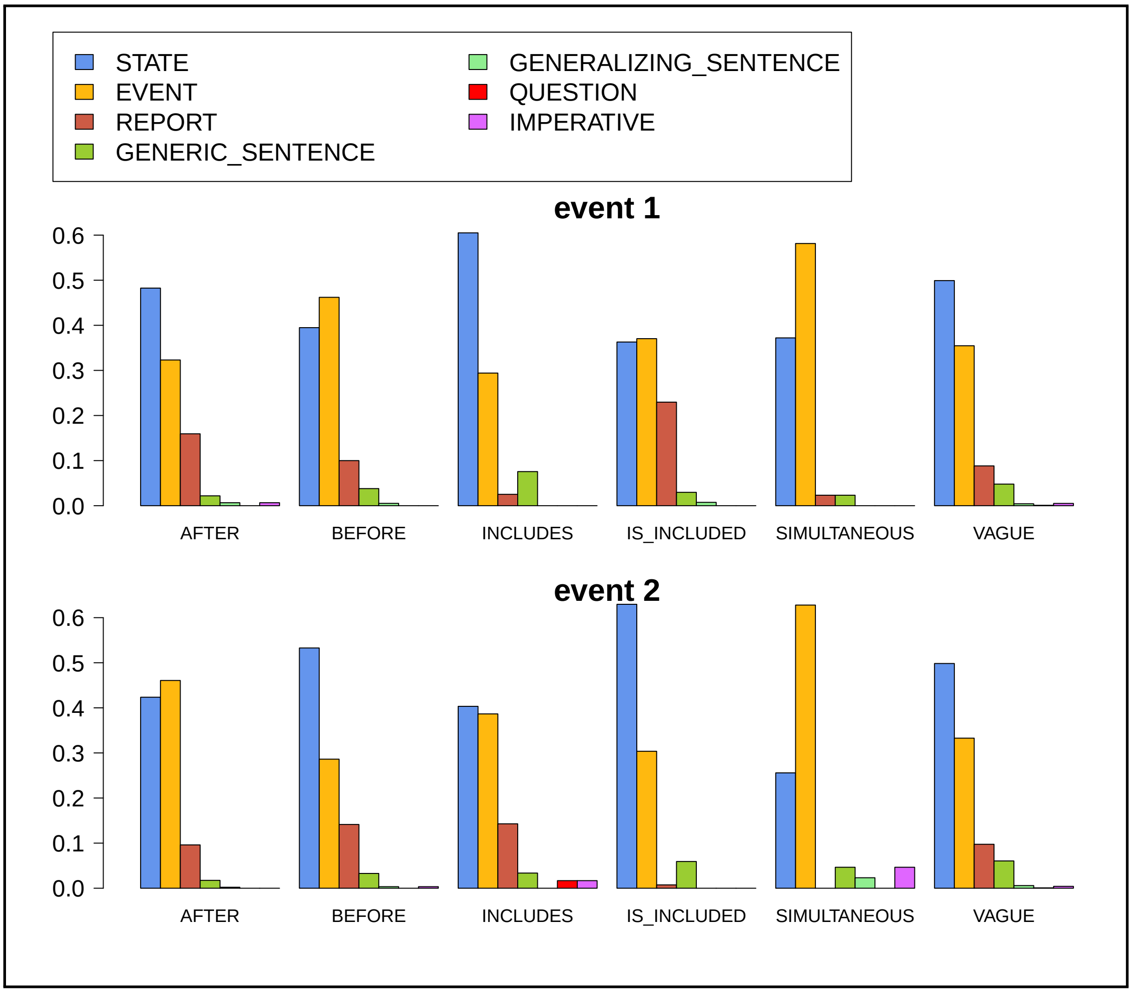

Figure 9.5: Situation entity types versus TimeML temporal relations (event-event relations).

In sum, this corpus study shows that situation entity types have more potential of being useful for automatic temporal relation identification than TimeML event classes, as they better capture differences between the various TimeML temporal relation types. We have conducted preliminary experiments on using situation entity type information in an automatic temporal relation identification system (Chambers et al., 2014). We did not achieve significant improvements over the very competitive rule-based and fine-tuned baseline system by integrating our predicted situation entity type information. We assume, however, that this is due to the limited size of the data set (there are only 1331 temporal relations in the test set with a skewed distribution of relations). Further research, especially a careful analysis of which sources of information need to be integrated in order to successfully leverage situation entity type information for temporal relation classification, is needed.

Chapter 10

Outlook

This chapter first discusses perspectives of using aspectual information for natural language processing. In the second part of this chapter, we describe concrete ideas for extensions of the work presented in this thesis.

10.1 Relevance of aspectual information for natural language processing

Aspect, as one of the sub-systems of language, crucially contributes to natural language understanding. This thesis explores the computational modeling of aspectual phenomena in written text. We expect that these models will provide a useful source of information for various research areas within natural language processing (NLP). We here discuss the potential of incorporating aspectual information into a variety of tasks or applications.

Temporal reasoning and discourse modes

The interpretation of the temporal structure of texts, i.e., inferring the temporal location and relative ordering of the situations mentioned, has been an important research area within computational linguistics for decades (Lascarides and Asher, 1993; Passonneau, 1988; Verhagen et al., 2007, 2010; UzZaman et al., 2013). It has been approached in a rule-based way and as a supervised machine learning task leveraging annotated corpora, as well as a combination of these two approaches (see, e.g., Pustejovsky et al., 2003b; Mani et al., 2006; D'Souza and Ng, 2013; Chambers et al., 2014).

Temporal location and relations are signaled by tense and aspect, both sub-systems of language (Comrie, 1976; Moens and Steedman, 1988). Tense locates a situation in the past, present or future relative to the time of speaking or writing, or the respective *reference time* (Reichenbach, 1947). Aspect, as illustrated in the introduction by example (1), repeated here for convenience, is a subjective category as the speaker or author can decide to make visible an entire situation or only part of it. Different choices of aspectual form cause different temporal interpretations of a discourse depending on the situations' types (Comrie, 1976; Smith, 1997, 2005; Siegel, 1998b).

(1) (a) The ship moved. (entire event)
(b) The ship was moving. (ongoing event / process)
(c) The ship was in motion. (state)

TimeML (Pustejovsky et al., 2003a), the current de-facto standard for temporal relation markup, labels each EVENT mention with a TimeML-defined *semantic event class*, which roughly corresponds to its situation type. As we have shown in Chapter 9, however, the semantic class of events does not seem to be strongly correlated with the various temporal relations in TimeBank (Pustejovsky et al., 2003b). However, as our preliminary corpus study on TimeBank in Chapter 9 demonstrates, the distribution of situation entity types (particularly States and Events) shows clear correlations with *before* and *after* relations.

Early computational work on temporal relation identification (Passonneau, 1988) computes temporal relations using tense and aspect, including situation type. This early work, however, consisted of a domain-specific Prolog-based system limited to sentence-internal relations. Costa and Branco (2012) successfully leverage aspect for temporal relation processing in Portuguese. For English, this has not yet been shown, but the results of our above mentioned case study are promising.

Most work on temporal relation identification to date has been conducted on different genres separately; news (UzZaman et al., 2013), clinical narratives (Styler IV et al., 2014; Bethard et al., 2016) and fables (Kolomiyets et al., 2012). But even within these genres, there is considerable variation with regard to how temporal relations are signaled. Specifically, each passage in a text is written in a particular mode of discourse in the sense of Smith (2003). Without claiming exhaustiveness, Smith lists the modes **Narrative**, **Report**, **Information**, **Description** and **Argument/Commentary**.

The **Narrative** mode (2) is used when telling stories, presenting predominantly States and Events that happen in a certain order. The narrative time moves forward as the text progresses.

(2) **Narrative**: I had called upon my friend, Mr. Sherlock Holmes, one day in the autumn of last year and found him in deep conversation with a very stout, florid-faced, elderly gentleman with fiery red hair. With an apology for my intrusion, I was about to withdraw when Holmes pulled me abruptly into the room and closed the door behind me.

In contrast, the situations in the **Report** mode (3) generally refer to the time of speaking or writing rather than being temporally located with respect to each other.

(3) **Report**: The protest started around 11:00 local time (UTC+1) on Whitehall opposite the police-guarded entrance to Downing Street, the Prime Minister's official residence. Just after 11:00, protesters blocked traffic on the northbound carriage in Whitehall. At 11:20, the police asked the protesters to move back on to the pavement, stating that they needed to balance the right to protest with the traffic building up. Around 11:29, the protest moved up Whitehall, past Trafalgar Square, along the Strand, passing by Aldwych and up Kingsway towards Holborn where the Conservative Party were holding their Spring Forum in the Grand Connaught Rooms hotel.[1]

In **Information** mode (see example (4)), it may not even be possible to align the situations temporally. Generic sentences and habituals form the background of a text and should not be directly linked into the story's temporal foreground.

(4) **Information**: The domestic dog (Canis lupus familiaris or Canis familiaris) is a domesticated canid which has been selectively bred for millennia for various behaviors, sensory capabilities, and physical attributes. Although initially thought to have originated as a manmade variant of an extant canid species (variously supposed as being the dhole, golden jackal, or gray wolf), extensive genetic studies undertaken during the 2010s indicate that dogs diverged from an extinct wolf-like canid in Eurasia 40,000 years ago.

We therefore argue that a successful domain- and genre-independent temporal relation processing system must be able to identify the discourse mode of a passage in order to correctly interpret the syntactic-semantic signals associated with the situations (Smith, 2005). One factor in recognizing a passage's discourse mode is determining the distribution of situation entity types in a passage, as this makes up part of the definition of discourse modes. For example, presence of many General Statives is an indicator for **Information** mode, while passages consisting mostly of States and Events are more likely in **Narrative** or **Report** mode.

Information extraction

Information extraction refers to the automatic extraction of structured information from unstructured resources such as natural language text (Sarawagi, 2008; Jiang, 2012). The aim of information extraction is to enable richer forms of queries for searching the original unstructured data. Temporal relation processing as presented above can in this sense be regarded as a sub-task within information extraction. In the natural language processing community, however, information extraction has mostly been associated with the identification of named entities and relationships among them.

[1] from Wikinews (`https://en.wikinews.org`): "Thousands march in London calling for David Cameron's resignation over tax affairs," April 9, 2016.

The most prominent information extraction program within natural language processing was the Automatic Contract Extraction (ACE) program, whose objective was to develop extraction technology for the automatic processing of natural language data.[2] Annotation of named entities in the ACE corpora includes assigning an *entity type* such as person (PER) or geo-political entity (GPE). In addition, *relations* are identified between pairs of named entities, e.g., `citizenship(Obama, USA)`. Most relevant to our work, named entities in ACE are classified as specific, generic, attributive, negatively quantified or underspecified (Walker et al., 2006). However, as we argue in Section 3.2.2, ACE's annotation scheme conflates the linguistic notions of genericity and specificity, where the former relates to kind-reference and the latter to whether the identity of a referent is known to the speaker or not. For accurate information extraction, a clean treatment of these two phenomena is certainly crucial. Distinguishing between statements about particular individuals or situations and generic sentences is an important part of human language understanding. Consider example (5): sentence (a) names characteristic attributes of a kind, which are inherent to every (typical) individual, and sentence (b) describes a specific individual.

(5) (a) <u>The modern domestic horse</u> has a life expectancy of 25 to 30 years. (***generic***)

(b) <u>Old Billy</u> lived to the age of 62. (***non-generic***)

The above example illustrates that generic and non-generic sentences differ substantially in their semantic impact and entailment properties. It can be inferred from sentence (5a) that a *typical* horse has a life expectancy of 25 to 30 years, and if we know that *Nelly* is a horse, we can infer that its life expectancy is 25 to 30 years. Sentence (5b) has no such properties, it only allows inferences about the particular individual *Old Billy*.

An automatic classifier that recognizes generic expressions would be extremely valuable for various kinds of information extraction systems that rely on "natural language understanding," including question answering systems which require textual entailment methods, and systems acquiring machine-readable knowledge from text (see also Van Durme, 2009). Machine-readable knowledge bases have different representations for statements corresponding to generic knowledge about kinds and knowledge about specific individuals. The non-generic sentence (5b) roughly speaking provides ABox content for a machine-readable knowledge base, i.e., knowledge about particular instances, e.g, "A is an instance of B / has property X." In contrast, the generic sentence (5a) feeds the TBox, i.e., knowledge of the form "All B are C." Reiter and Frank (2010) provide a detailed discussion of the relevance of the distinction between classes and instances for automatic ontology construction.

Factuality recognition

Natural language understanding involves interpreting whether an event mentioned in a text actually happened, or whether the speaker or writer believes that it happened

[2] `https://www.ldc.upenn.edu/collaborations/past-projects/ace`

(de Marneffe et al., 2012). Proper recognition of factuality has applications in temporal reasoning (as already explained above), as well as in information extraction, question answering or summarization tasks (see also Lotan et al., 2013; Lee et al., 2015). We illustrate this using the following example. The first sentence in example (6) is habitual: it generalizes over situations in which Bill normally drinks coffee.

(6) Bill usually drinks coffee after lunch. Yesterday, however, he did not get any and as a result he was ill-tempered all afternoon.

Habituals are known to allow for exceptions; the first sentence of (6) is considered to be true even if Bill occasionally does not get his coffee (Carlson, 2005). A system unable to detect that this sentence is habitual might extract `drink(Bill, coffee)` as an event representation from the above text. The text, however, focuses on a particular day on which Bill did not drink coffee, and the habitual sentence provides background knowledge rather than information about a particular event. Precise event extraction and representation therefore requires the modeling of aspectual distinctions such as habituality.

Other potential applications

Aspectual classification also has the potential to improve **machine translation** systems (Siegel, 1998b; Dorr, 1991): for example, when translating from a language that does not have perfective markers (such as English) to a language that does have them (such as Bulgarian or Russian), an aspectual categorization of the input may improve the system's output. A similar idea for choosing the appropriate tense when translating from Chinese to English has been explored by Zhang and Xue (2014).

Another relevant field of application is **coreference resolution**: there is no consensus in the community about how or whether to annotate coreference for mentions referring to kinds (Nedoluzhko, 2013). Generic mentions are often not linked into the coreference structures in annotated corpora, which inevitably leads to problems during testing time. Reliable methods for identifying generic mentions could help automatic coreference resolution systems to treat such cases in a more principled way.

10.2 Next steps

Above, we have given the broad picture of how natural language processing could profit from integrating aspectual information. In this section, we discuss several ideas on how to improve upon the representation and learning of aspectual information as presented in this thesis.

Segmentation. We have based the segmentation of texts into situation segments on an existing discourse parser (Soricut and Marcu, 2003), and given reasons for why this is a good fit for the research endeavor presented in this thesis in Chapter 4. A valid question is whether we could have simply used verbs as the unit for annotation instead, i.e., any token carrying a verbal part-of-speech tag. In fact, in future work, we plan to present text in the original formatting rather than one segment per line, marking verbs as the targets for situation entity annotation. This, however, does not avoid the question of situation segmentation as we do not consider all verbs to introduce situation entities to the discourse as discussed in Chapter 4: we will either have to put this task on the annotator, or filter out verbs that we do not consider to invoke situation entities before marking targets in a way similar to what we do now. We have described a preliminary attempt at creating such a segmenter in Chapter 9, however, this method needs a more careful evaluation and a more principled way of determining which constituents belong to the clause of a particular main verb.

The second open question regarding segmentation regards the treatment of nominal constructions. While it is not trivial to determine which cases introduce situation entities to the discourse, some of them – such as (7), repeated from Chapter 4 here for convenience – clearly do.

(7) The destruction of the old town hall really was a big loss for our city.

To exhaustively model the temporal structure of a text, we believe that we should take some of these constructions into account as invoking situation entities. More research into how well humans agree on identifying such cases is necessary. A starting point is the analysis of parallel corpora: nominal constructions that are translated as verbs are good candidates for situation-invoking nominal constructions.

Subtyping of State. The situation entity type State covers many different types of clauses: situations that attribute properties to an individual, events reported in the Perfect, as well as clauses containing negations or modal verbs.[3] From an application point of view, it may be desirable to distinguish between different subtypes of State.

Our label **static** covers uses of lexically stative predicates that do not generalize over situations. In future work, it may be very useful to distinguish individual- from stage-level predicates (Carlson, 1977a). Individual-level stative sentences like (8) describe properties which hold as long as the individual that they are attributed to exists. In contrast, stage-level predicates are used for so-called *episodic statives* as in (9), they describe states which only hold for a limited period.

(8) My cat is black. (**static**)

(9) My cat is hungry. (**static**)

This distinction is for example relevant for information extraction: while the information in (8) is of the kind that would be entered into a knowledge base, the information in (9) is more relevant to the (e.g., temporal) representation of a story. Near-term future

[3] We did some preliminary experiments breaking up the State class according to these phenomena, but this did not increase classification accuracy for situation entity types.

work includes an annotation study for determining how well humans agree on this distinction, and the evaluation of how well linguistic indicators similar to those described in Section 7.1.6 capture the difference.

Another related research direction is the classification of the different types of ***static*** clauses, e.g., the different senses of modality (Ruppenhofer and Rehbein, 2012; Zhou et al., 2015; Marasović et al., 2016). These senses include, among others, the distinction between *dynamic* (10a) and *epistemic* (10b) modality.[4]

(10) My dad could have done it.
(a) My dad had the ability/capacity to do it. (**dynamic**)
(b) It's possible that my dad did it. (**epistemic**)

As discussed in Section 5.1.4, modality, negation and habituality interact at the clausal level. For the purpose of creating the corpus presented in this thesis, we have decided to mark clauses with modal verbs as habitual if they imply that something happened regularly. A deeper investigation of the interaction of modal senses with clause-level aspect, both on a theoretical and a practical level, is needed.

Subtyping Event. We here discuss two properties of Events that we do not model yet, but which we plan to address in future work. The first is related to inherent lexical aspect and regards the modeling of whether the event type includes natural endpoints, i.e., whether it is telic (see Section 2.1.2). The second distinction is related to grammatical aspect, more specifically, whether an eventuality is presented as bounded or unbounded in context. It has been observed that events that are presented as ongoing share some linguistic characteristics with States, e.g., they do not move the narrative time. Recognizing boundedness is relatively easy in English (Loáiciga and Grisot, 2016, see Section 3.1.5) as the Progressive is a strong indicator. In other languages, such as Chinese, it is more difficult as it depends on many contextual and pragmatic factors (Xue and Zhang, 2014). We envision a joint treatment of these two related phenomena, i.e., telicity and boundedness. It is on our research agenda to investigate to what extent parallel data from various languages that encode part of the distinctions morphologically can be leveraged to create cross-lingual models of fine-grained eventuality types.

Leveraging uncertainty in the corpus annotation. In the experiments presented in this thesis, we have worked with a single gold standard created from the various annotators' labels via majority voting. As we have preserved the annotations of each annotator as well, the corpus offers the possibility for further study of easy and difficult cases, not only in a corpus-linguistic, but also in a computational way. The set of labels given to each item provides information about the degree of certainty with which the item can be assigned to the respective categories. In near-future work, we plan to leverage this information during training and evaluation of our computational models, following ideas of Plank et al. (2014).

[4] Example from Ruppenhofer and Rehbein (2012). *Dynamic* modality indicates the modal sense of ability, i.e., this term has nothing in common with our use of ***dynamic*** in the context of lexical aspect.

Chapter 11

Conclusion

This thesis work addresses the manual and automatic identification of situation entity types following the theoretical work by Smith (2003). The key contributions of this thesis are:

- We have conducted a large-scale in-depth **corpus study** on labeling texts from a variety of genres with clause-level aspectual and semantic information. We have found that our annotators substantially agree on lexical aspectual class, habituality, genericity and situation entity type. The resulting corpus also forms the basis for training and evaluating machine learning approaches to classifying the above mentioned phenomena.
- We have created **computational models** for all aspectual distinctions that are annotated in our corpus, and evaluated them on held-out test sets. We have related the automatic approaches for each of the classification tasks to the respective previous work and outperformed the prior approaches in each case. Our publicly available system can readily be applied to any written English text, making it easy to explore the utility of situation entity types for other natural language processing tasks.

Our annotation scheme and guidelines for annotating situation entity types (Friedrich and Palmer, 2014b) follow established traditions in linguistics and semantic theory. When applying these to a large number of natural texts, though, we came across a number of borderline cases where it is not easy to select just one situation entity type label. The most difficult case is the identification of Generic Sentences, which are defined as making a statement about a kind or class. The difficulty of this decision varies with the genre and domain being annotated – judging genericity in encyclopedic entries about animals or botany turned out to be relatively easy, while the task proved inherently difficult for genres discussing more abstract concepts, such as journal articles or argumentative essays (see also Becker et al., 2016).

Our study and modeling of situation entity types combines a variety of linguistic phenomena at the interface of syntax and semantics, and thus opens up research opportunities in many directions. As detailed in Chapter 10, aspectual information is highly relevant to temporal reasoning, factuality recognition and machine translation. Identifying noun phrases that refer to kinds is a step towards more accurate information extraction or coreference resolution methods. The linguistic phenomena modeled in this thesis, in addition to identifying the type of a given situation, capture distinctions which reflect how the writer or speaker has chosen to represent that situation in discourse. Modeling this part of the meaning of a text brings us a step closer to natural language understanding.

List of Figures

List of Tables

Bibliography

Anna Aalstein and Patrick Blackburn. An aspectual classification of Polish verbs. *Unpublished manuscript*, 2007.

Steven P Abney. Parsing by chunks. In *Principle-based parsing*, pages 257–278. Springer, 1991.

Jean-Michel Adam. *Les textes: types et prototypes: récit, description, argumentation, explication et dialogue*. Armand Colin, 2011.

James F Allen. An interval-based representation of temporal knowledge. In *IJCAI*, volume 81, pages 221–226, 1981.

Ron Artstein and Massimo Poesio. Bias decreases in proportion to the number of annotators. In *Proceedings of FG-MoL, The 10th conference on Formal Grammar and The 9th Meeting on Mathematics of Language*, pages 141–150, Edinburgh, UK, August 2005.

Ron Artstein and Massimo Poesio. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, 2008.

Nicholas Asher. *Reference to Abstract Objects in Discourse: A Philosophical Semantics for Natural Language Metaphysics*, volume 50 of *Studies in Linguistics and Philosophy*. Springer Science & Business Media, 1993.

Nicholas Asher and Michael Morreau. What Some Generic Sentences Mean. In Gregory N. Carlson and Francis Jeffry Pelletier, editors, *The Generic Book*, Studies in Communication, Media, and Public Opinion, pages 300–338. University Of Chicago Press, 1995.

Guy Aston and Lou Burnard. *The BNC handbook: exploring the British National Corpus with SARA*. Capstone, 1998.

John Langshaw Austin. *How to do things with words*, volume 1955. Oxford university press, 1975.

Harald R. Baayen, Richard Piepenbrock, and Leon Gulikers. CELEX2. Philadelphia: Linguistic Data Consortium, 1996.

Emmon Bach. The algebra of events. *Linguistics and philosophy*, 9(1):5–16, 1986.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. Abstract Meaning Representation for Sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria, August 2013.

Maria Becker, Alexis Palmer, and Anette Frank. Argumentative texts and clause types. In *Proceedings of the 3rd Workshop on Argument Mining*, Berlin, Germany, August 2016.

Cosmin Adrian Bejan and Sanda Harabagiu. Unsupervised event coreference resolution with rich linguistic features. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1412–1422, Uppsala, Sweden, July 2010.

Michael Bennett and Barbara H Partee. *Toward the logic of tense and aspect in English*. Blackwell Publishing Ltd., Wiley Online Library, 1978.

Steven Bethard. ClearTK-TimeML: A minimalist approach to TempEval 2013. In *Second Joint Conference on Lexical and Computational Semantics (* SEM)*, volume 2, pages 10–14, Atlanta, Georgia, June 2013.

Steven Bethard and James H Martin. Identification of event mentions and their semantic class. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 146–154, Sydney, Australia, July 2006.

Steven Bethard, Guergana Savova, Wei-Te Chen, Leon Derczynski, James Pustejovsky, and Marc Verhagen. SemEval-2016 Task 12: Clinical TempEval. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1052–1062, San Diego, California, June 2016.

Archna Bhatia, Chu-Cheng Lin, Nathan Schneider, Yulia Tsvetkov, Fatima Talib Al-Raisi, Laleh Roostapour, Jordan Bender, Abhimanu Kumar, Lori Levin, Mandy Simons, and Chris Dyer. Automatic classification of communicative functions of definiteness. In *Proceedings of the 25th International Conference on Computational Linguistics (Coling)*, pages 1059–1070, Dublin, Ireland, August 2014a.

Archna Bhatia, Mandy Simons, Lori Levin, Yulia Tsvetkov, Chris Dyer, and Jordan Bender. A unified annotation scheme for the semantic/pragmatic components of definiteness. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, Rejkjavik, Iceland, May 2014b.

Ann Bies, Mark Ferguson, Karen Katz, Robert MacIntyre, Victoria Tredinnick, Grace Kim, Mary Ann Marcinkiewicz, and Britta Schasberger. Bracketing guidelines for Treebank II style Penn Treebank project. *University of Pennsylvania*, 97:100, 1995.

Anders Björkelund, Kerstin Eckart, Arndt Riester, Nadja Schauffler, and Katrin Schweitzer. The extended DIRNDL corpus as a resource for automatic coreference and bridging resolution. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, pages 3222–3228, Rejkjavik, Iceland, May 2014.

Kristina Nilsson Björkenstam. SUC-CORE: A Balanced Corpus Annotated with Noun Phrase Coreference. *Northern European Journal of Language Technology (NEJLT)*, 3: 19–39, 2013.

Alena Böhmová, Jan Hajič, Eva Hajičová, and Barbora Hladká. The Prague dependency treebank. In *Treebanks*, pages 103–127. Springer, 2003.

Nora Boneh and Edit Doron. Hab and Gen in the Expression of Habituality. In Alda Mari, Claire Beyssade, and Fabio Del Prete, editors, *Genericity*, volume 43, page 176. Oxford University Press, 2013.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Michael R Brent. Automatic semantic classification of verbs from their syntactic contexts: an implemented classifier for stativity. In *Proceedings of the fifth conference on European chapter of the Association for Computational Linguistics (ACL)*, pages 222–226, Berlin, Germany, April 1991.

Jean-Paul Bronckart. *Activité langagière, textes et discours. Pour un interactionisme socio-discursif.* Delachaux et Niestlé, 1997.

Peter F Brown, Peter V Desouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. Class-based n-gram models of natural language. *Computational linguistics*, 18(4): 467–479, 1992.

Bianka Buschbeck, Renate Henschel, Iris Höser, Gerda Klimonow, Andreas Küstner, and Ingrid Starke. Limits of a sentence based procedural approach for aspect choice in German-Russian MT. In *Proceedings of the fifth conference on European chapter of the Association for Computational Linguistics (ACL)*, pages 269–274, 1991.

Miriam Butt, Helge Dyvik, Tracy Holloway King, Hiroshi Masuichi, and Christian Rohrer. The parallel grammar project. In *Proceedings of the 2002 workshop on Grammar engineering and evaluation*, volume 15, pages 1–7, Taipei, Taiwan, August 2002.

Defang Cao, Wenjie Li, Chunfa Yuan, and Kam-Fai Wong. Automatic Chinese aspectual classification using linguistic indicators. *International Journal of Information Technology*, 12(4):99–109, 2006.

Gregory N Carlson. *Reference to kinds in English*. PhD thesis, Amherst, Massachusetts, 1977a.

Gregory N Carlson. A unified analysis of the English bare plural. *Linguistics and philosophy*, 1(3):413–457, 1977b.

Gregory N Carlson. Truth-Conditions of Generic Sentences: Two Contrasting Views. In Gregory N. Carlson and Francis Jeffry Pelletier, editors, *The Generic Book*, Studies in Communication, Media, and Public Opinion, pages 224–237. University Of Chicago Press, 1995.

Gregory N Carlson. Generics, habituals and iteratives. In Keith Brown, editor, *The Encyclopedia of Language and Linguistics*. Elsevier Ltd, 2005.

Lynn Carlson and Daniel Marcu. Discourse tagging reference manual. *ISI Technical Report ISI-TR-545*, 54, 2001.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. Building a Discourse-tagged Corpus in the Framework of Rhetorical Structure Theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue - Volume 16*, SIGDIAL '01, pages 1–10, Stroudsburg, PA, USA, 2001.

Taylor Cassidy, Bill McDowell, Nathanel Chambers, and Steven Bethard. An annotation framework for dense event ordering. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, Baltimore, USA, June 2014.

Nathanael Chambers. Navytime: Event and time ordering from raw text. Technical report, DTIC Document, 2013.

Nathanael Chambers, Taylor Cassidy, Bill McDowell, and Steven Bethard. Dense event ordering with a multi-pass architecture. *Transactions of the Association for Computational Linguistics (TACL)*, 2:273–284, 2014.

Eugene Charniak. A Maximum-entropy-inspired Parser. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference (NAACL)*, pages 132–139, Seattle, Washington, April 2000.

Christelle Cocco. Discourse type clustering using pos n-gram profiles and high-dimensional embeddings. In *Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 55–63, 2012.

Christelle Cocco, Raphaël Pittier, François Bavaud, and Aris Xanthos. Segmentation and clustering of textual sequences: a typological approach. In *Proceedings of Recent Advances in Natural Language Processing (RANLP)*, pages 427–433, Hissar, Bulgaria, September 2011.

Jacob Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213, 1968.

Bernard Comrie. *Aspect: An introduction to the study of verbal aspect and related problems*, volume 2 of *Cambridge Textbooks in Linguistics*. Cambridge University Press, 1976.

Corinna Cortes and Vladimir Vapnik. Support-Vector Networks. *Machine Learning*, 20 (3):273–297, September 1995. ISSN 0885-6125.

Francisco Costa and António Branco. Aspectual type and temporal relation classification. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Avignon, France, April 2012.

William Croft, Pavlina Peskova, and Michael Regan. Annotation of causal and aspectual structure of events in RED: a preliminary report. In *Proceedings of the Fourth Workshop on Events*, pages 8–17, San Diego, California, June 2016.

James Curran, Stephen Clark, and Johan Bos. Linguistically Motivated Large-Scale NLP with C&C and Boxer. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Companion Volume Proceedings of the Demo and Poster Sessions*, pages 33–36, Prague, Czech Republic, June 2007.

Agata Cybulska and Piek Vossen. Using a sledgehammer to crack a nut? Lexical diversity and event coreference resolution. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, pages 4545–4552, Rejkjavik, Iceland, May 2014.

Östen Dahl. The Marking of the episodic/generic distinction in tense-aspect systems. In Gregory N. Carlson and Francis Jeffry Pelletier, editors, *The Generic Book*, Studies in Communication, Media, and Public Opinion, pages 412–425. University Of Chicago Press, 1995.

Marie-Catherine de Marneffe and Christopher D Manning. The Stanford typed dependencies representation. In *Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pages 1–8, Manchester, UK, August 2008a.

Marie-Catherine de Marneffe and Christopher D Manning. Stanford typed dependencies manual. Technical report, Stanford University, 2008b.

Marie-Catherine de Marneffe, Christopher D Manning, and Christopher Potts. Did it happen? The pragmatic complexity of veridicality assessment. *Computational linguistics*, 38(2):301–333, 2012.

Leon Derczynski and Robert Gaizauskas. Temporal Relation Classification using a Model of Tense and Aspect. In *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP)*, pages 118–122, Hissar, Bulgaria, September 2015.

Barbara Di Eugenio and Michael Glass. The kappa statistic: A second look. *Computational linguistics*, 30(1):95–101, 2004.

George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie Strassel, and Ralph M Weischedel. The Automatic Content Extraction (ACE) Program-Tasks, Data, and Evaluation. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, volume 2, page 1, Lisbon, Portugal, May 2004.

Bonnie J Dorr. A two-level knowledge representation for machine translation: lexical semantics and tense/aspect. In *Workshop of SIGLEX (Special Interest Group within ACL on the Lexicon)*, pages 269–287. Springer, 1991.

Bonnie J Dorr. Large-scale acquisition of LCS-based lexicons for foreign language tutoring. In *Proceedings of the fifth conference on Applied natural language processing*, pages 139–146, Washington, D.C., March 1997.

Bonnie J Dorr. LCS Verb Database, Online Software Database of Lexical Conceptual Structures and Documentation. University of Maryland, 2001.

Bonnie J Dorr and Mari Broman Olsen. Deriving verbal and compositional lexical aspect for NLP applications. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics (EACL)*, pages 151–158, Madrid, Spain, 1997.

David R Dowty. *Word meaning and Montague grammar – The Semantics of Verbs and Times in Generative Semantics and in Montague's PTQ*. Studies in Linguistics and Philosophy. 1979.

David R Dowty. The effects of aspectual class on the temporal structure of discourse: semantics or pragmatics? *Linguistics and philosophy*, 9(1):37–61, 1986.

Jennifer D'Souza and Vincent Ng. Classifying Temporal Relations with Rich Linguistic Knowledge. In *The 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 918–927, Atlanta, Georgia, June 2013.

Richard Eckart de Castilho and Iryna Gurevych. A broad-coverage collection of portable NLP components for building shareable analysis pipelines. In *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT*, pages 1–11, Dublin, Ireland, August 2014.

Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, Massachussetts, 1998.

David Ferrucci and Adam Lally. UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 10(3-4):327–348, 2004.

Hana Filip. *Aspect, eventuality types and nominal reference*. Taylor & Francis, 1999.

Hana Filip. Lexical aspect. In Robert I Binnick, editor, *The Oxford handbook of tense and aspect*, pages 721–751. Oxford University Press, 2012.

Hana Filip and Gregory N Carlson. Sui generis genericity. *University of Pennsylvania Working Papers in Linguistics*, 4(2):7, 1997.

Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

W Nelson Francis and Henry Kučera. Brown corpus manual. *Brown University*, 1979. http://clu.uni.no/icame/brown/bcm.html.

Aviva Freedman and Peter Medway. Locating genre studies: Antecedents and prospects. *Genre and the new rhetoric*, pages 1–20, 1994.

Annemarie Friedrich and Alexis Palmer. Automatic prediction of aspectual class of verbs in context. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, Baltimore, USA, June 2014a.

Annemarie Friedrich and Alexis Palmer. Situation entity annotation. In *Proceedings of the 8th Linguistic Annotation Workshop (LAW VIII)*, Dublin, Ireland, August 2014b.

Annemarie Friedrich and Manfred Pinkal. Discourse-sensitive Automatic Identification of Generic Expressions. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL)*, Beijing, China, July 2015a.

Annemarie Friedrich and Manfred Pinkal. Automatic recognition of habituals: a three-way classification of clausal aspect. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Lisbon, Portugal, September 2015b.

Annemarie Friedrich, Kleio-Isidora Mavridou, and Alexis Palmer. *Situation Entity Types – Annotation Manual, version 1.1.* Saarland University, April 2015a.

Annemarie Friedrich, Alexis Palmer, Melissa Peate Sørensen, and Manfred Pinkal. Annotating genericity: a survey, a scheme, and a corpus. In *Proceedings of the 9th Linguistic Annotation Workshop (LAW IX)*, Denver, Colorado, USA, June 2015b.

Annemarie Friedrich, Alexis Palmer, and Manfred Pinkal. Situation entity types: automatic classification of clause-level aspect. In *In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, Berlin, Germany, August 2016.

Howard B Garey. Verbal aspect in French. *Language*, 33(2):91–110, 1957.

Andrew Garrett. On the origin of auxiliary do. *English Language and Linguistics*, 2(02): 283–330, 1998.

Barbara Gawronska. Aspect - a Problem for MT. In *Proceedings of The 15th International Conference on Computational Linguistics (Coling)*, Nantes, France, August 1992.

Susan A Gelman and Twila Tardif. A cross-linguistic comparison of generic noun phrases in English and Mandarin. *Cognition*, 66(3):215–248, 1998.

David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. English Gigaword. *Linguistic Data Consortium, Philadelphia*, 2003.

Ralph Grishman and Beth Sundheim. Message understanding conference-6: A brief history. In *Proceedings of the 16th International Conference on Computational Linguistics (Coling)*, pages 466–471, Copenhagen, Denmark, August 1996.

Yufan Guo, Anna Korhonen, Maria Liakata, Ilona Silins Karolinska, Lin Sun, and Ulla Stenius. Identifying the information structure of scientific abstracts: an investigation of three different schemes. In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, pages 99–107, Uppsala, Sweden, July 2010.

Yufan Guo, Anna Korhonen, and Thierry Poibeau. A weakly-supervised approach to argumentative zoning of scientific documents. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 273–283, Edinburgh, UK, July 2011.

Valentine Hacquard. On the interaction of aspect and modal auxiliaries. *Linguistics and Philosophy*, 32(3):279–315, 2009.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.

Aurelie Herbelot and Ann Copestake. Annotating genericity: How do humans decide? (A case study in ontology extraction). *Studies in Generative Grammar 101*, page 103, 2009.

Aurelie Herbelot and Ann Copestake. Annotating underquantification. In *Proceedings of the Fourth Linguistic Annotation Workshop (LAW-IV)*, pages 73–81, Uppsala, Sweden, July 2010.

Aurelie Herbelot and Ann Copestake. Formalising and specifying underquantification. In *Proceedings of the Ninth International Conference on Computational Semantics (IWCS)*, pages 165–174, Oxford, UK, January 2011.

Jürgen Hermes, Michael Richter, and Claes Neuefeind. Automatic Induction of German Aspectual Verb Classes in a Distributional Framework. In *Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology (GSCL)*, pages 122–129, Duisburg-Essen, Germany, October 2015.

Erhard Hinrichs. *A compositional semantics for Aktionsarten and NP reference in English*. PhD thesis, The Ohio State University, 1985.

Erhard Hinrichs, Sandra Kübler, Karin Naumann, Heike Telljohann, and Julia Trushkina. Recent developments in linguistic annotations of the TüBa-D/Z treebank. In *Proceedings of the Third Workshop on Treebanks and Linguistic Theories*, pages 51–62, 2004.

Jens Holt. *Etudes d'aspect*. Universitetsforlaget i Aarhus ejnar munksgaard, 1943.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. Ontonotes: the 90% solution. In *Proceedings of Human Language Technology conference - North American chapter of the Association for Computational Linguistics annual meeting (HLT-NAACL)*, pages 57–60, New York City, New York, June 2006.

Nancy Ide and Catherine Macleod. The American National Corpus: A standardized resource of American English. In *Proceedings of Corpus Linguistics 2001*, volume 3, Lancaster, UK, March 2001.

Nancy Ide and Keith Suderman. The American National Corpus First Release. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, Lisbon, Portugal, May 2004.

Nancy Ide, Collin Baker, Christiane Fellbaum, and Charles Fillmore. MASC: The manually annotated sub-corpus of American English. In *In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco, 2008.

Nancy Ide, Christiane Fellbaum, Collin Baker, and Rebecca Passonneau. The manually annotated sub-corpus: A community resource for and by the people. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 68–73, Uppsala, Sweden, July 2010.

Rei Ikuta, William F Styler IV, Mariah Hamang, Tim O'Gorman, and Martha Palmer. Challenges of adding causation to richer event descriptions. In *Proceedings of the 2nd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, page 12, Baltimore, Maryland, June 2014.

Jing Jiang. Information extraction from text. In *Mining text data*, pages 11–41. Springer, 2012.

Eric Joanis, Suzanne Stevenson, and David James. A general feature space for automatic verb classification. *Natural Language Engineering*, 14(03):337–367, 2007.

Hyuckchul Jung and Amanda Stent. Att1: Temporal annotation using big windows and rich syntactic and semantic features. In *Second Joint Conference on Lexical and Computational Semantics (* SEM)*, volume 2, pages 20–24, 2013.

Hans Kamp and Christian Rohrer. A discourse representation theory account of tense in French. *Unpublished manuscript*, 1989.

Graham Katz. On the stativity of the English perfect. *Perfect explorations*, pages 205–234, 2003.

Richard Keelan. Lexical aspectual classification. Master's thesis, University of Ottawa, 2012.

Anthony Kenny. *Action, emotion and will.* Routledge, 1963.

Judith L Klavans and Martin Chodorow. Degrees of stativity: the lexical representation of verb aspect. In *Proceedings of the 14th conference on Computational linguistics (Coling)*, volume 4, pages 1126–1131, Nantes, France, August 1992.

Beata Beigman Klebanov and Eyal Beigman. From annotator agreement to noise models. *Computational Linguistics*, 35(4):495–503, 2009.

Dan Klein and Christopher D Manning. Fast exact inference with a factored model for natural language parsing. In *Advances in neural information processing systems*, pages 3–10, 2002.

Bryan Klimt and Yiming Yang. The enron corpus: A new dataset for email classification research. In *Machine learning: ECML 2004*, pages 217–226. Springer, 2004.

Roman Klinger and Christoph M Friedrich. Feature subset selection in conditional random fields for named entity recognition. In *Proceedings of Recent Advances in Natural Language Processing (RANLP)*, 2009.

Roman Klinger and Katrin Tomanek. Classical probabilistic models and conditional random fields. *TU Dortmund Algorithm Engineering Report*, December 2007.

Oleksandr Kolomiyets, Steven Bethard, and Marie-Francine Moens. Extracting Narrative Timelines As Temporal Dependency Structures. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 88–97, Jeju Island, Korea, July 2012.

Manfred Krifka and Claudia Gerstner. An outline of genericity. In *Seminar für natürlichsprachliche Systeme, University of Tübingen*, number SNS-Bericht 87-25, 1987.

Manfred Krifka, Francis Jeffrey Pelletier, Gregory N. Carlson, Alice ter Meulen, Godehard Link, and Gennaro Chierchia. Genericity: An Introduction. In Gregory N. Carlson and Francis Jeffry Pelletier, editors, *The Generic Book*, Studies in Communication, Media, and Public Opinion, pages 1–124. University Of Chicago Press, 1995.

Klaus Krippendorff. *Content analysis: An introduction to its methodology.* Sage, 1980.

Anna Kupść. Two approaches to aspect assignment in an English-Polish machine translation system. In *Proceedings of the 7th International EAMT workshop on MT and other Language Technology Tools, Improving MT through other Language Technology Tools: Resources and Tools for Building MT*, pages 17–24, 2003.

John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the eighteenth International Conference on Machine Learning (ICML)*, volume 1, pages 282–289, Williamstown, Massachussetts, 2001.

J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174, 1977.

Maria Lapata and Chris Brew. Using subcategorization to resolve verb class ambiguity. In *Proceedings of Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP)*, pages 266–274, Hong Kong, October 1999.

Alex Lascarides. The Progressive and the Imperfective Paradox. *Synthese*, 87(3):401–447, 1991.

Alex Lascarides and Nicholas Asher. Temporal interpretation, discourse relations and commonsense entailment. *Linguistics and Philosophy*, 16(5):437–493, 1993.

John Michael Lawler. *Studies in English generics*. PhD thesis, The University of Michigan, 1973.

Kenton Lee, Yoav Artzi, Yejin Choiy, and Luke Zettlemoyer. Event detection and factuality assessment with non-expert supervision. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Lisbon, Portugal, 2015.

Geoffrey N Leech. *Meaning and the English verb*. Pearson Education, 1971.

Beth Levin. *English Verb Classes and Alternations – A preliminary investigation*. University of Chicago Press, Chicago, Illinois, 1993.

Jianguo Li and Chris Brew. Which Are the Best Features for Automatic Verb Classification. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 434–442, Columbus, Ohio, June 2008.

Percy Liang. *Semi-supervised learning for natural language*. PhD thesis, Massachusetts Institute of Technology, Cambridge, Massachussetts, 2005.

Godehard Link. Generic information and dependent generics. In Gregory N. Carlson and Francis Jeffry Pelletier, editors, *The Generic Book*, Studies in Communication, Media, and Public Opinion, pages 358–382. University Of Chicago Press, 1995.

Hector Llorens, Estela Saquete, and Borja Navarro-Colorado. TimeML events recognition and classification: learning CRF models with semantic roles. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling)*, pages 725–733, Beijing, China, August 2010.

Sharid Loáiciga and Cristina Grisot. Predicting and using a pragmatic component of lexical aspect. *Linguistic Issues in Language Technology, Special issue on Modality in Natural Language Understanding*, 13, 2016.

Sharid Loaiciga, Thomas Meyer, and Andrei Popescu-Belis. English-French Verb Phrase Alignment in Europarl for Tense Translation Modeling. In *The Ninth Language Resources and Evaluation Conference (LREC)*, Rejkjavik, Iceland, May 2014.

Amnon Lotan, Asher Stern, and Ido Dagan. TruthTeller: Annotating Predicate Truth. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 752–757, Atlanta, Georgia, June 2013.

Annie Louis and Ani Nenkova. Automatic identification of general and specific sentences by leveraging discourse annotations. In *Proceedings of The 5th International Joint Conference on Natural Language Processing (IJCNLP)*, pages 605–613, Chiang Mai, Thailand, November 2011.

Annie Louis and Ani Nenkova. A corpus of general and specific sentences from news. In *Proceedings of The eighth international conference on Language Resources and Evaluation (LREC)*, pages 1818–1821, Istanbul, Turkey, May 2012.

David M Magerman. Statistical decision-tree models for parsing. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics (ACL)*, pages 276–283, Cambridge, Massachusetts, June 1995.

Inderjeet Mani, Marc Verhagen, Ben Wellner, Chong Min Lee, and James Pustejovsky. Machine learning of temporal relations. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics (ACL)*, pages 753–760, Sydney, NSW, Australia, July 2006.

William C Mann and Sandra A Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281, 1988.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J Bethard, and David McClosky. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of the 52nd annual meeting of the Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60, Baltimore, Maryland, June 2014.

Ana Marasović, Mengfei Zhou, Alexis Palmer, and Anette Frank. Modal Sense Classification At Large: Paraphrase-Driven Sense Projection, Semantically Enriched Classification Models and Cross-Genre Evaluations. In *Linguistic Issues in Language Technology, Special issue on Modality in Natural Language Understanding*, volume 14 (2), Stanford, CA., 2016. CSLI Publications.

Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a large annotated corpus of English: The Penn Treebank. *Computational linguistics*, 19(2): 313–330, 1993.

Thomas A. Mathew. Supervised categorization for habitual versus episodic sentences. Master's thesis, Faculty of the Graduate School of Arts and Sciences of Georgetown University, 2009.

Thomas A. Mathew and E. Graham Katz. Supervised Categorization of Habitual and Episodic Sentences. In *Sixth Midwest Computational Linguistics Colloquium*, Bloomington, Indiana: Indiana University, 2009.

Kleio-Isidora Mavridou, Annemarie Friedrich, Melissa Peate Sorensen, Alexis Palmer, and Manfred Pinkal. Linking discourse modes and situation entities in a cross-linguistic corpus study. In *Proceedings of Linking Models of Lexical, Sentential and Discourse-level Semantics (LSDSem 2015)*, Lisbon, Portugal, September 2015.

Andrew K McCallum. MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu, 2002.

Michael C. McCord. Slot grammar: A system for simpler construction of practical natural language grammars. In *Natural Language and Logic: International Scientific Symposium Hamburg*, Lecture Notes in Computer Science 459, pages 118–145, Berlin, May 1990. Springer Verlag.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

Paola Merlo and Suzanne Stevenson. Automatic verb classification based on statistical distributions of argument structure. *Computational Linguistics*, 27(3):373–408, 2001.

Christian M. Meyer, Margot Mieskes, Christian Stab, and Iryna Gurevych. DKPro Agreement: An Open-Source Java Library for Measuring Inter-Rater Agreement. In *Proceedings of the 25th International Conference on Computational Linguistics: System Demonstrations (Coling)*, pages 105–109, Dublin, Ireland, August 2014.

Alexis Mitchell, Stephanie Strassel, Mark Przybocki, JK Davis, George Doddington, Ralph Grishman, Adam Meyers, Ada Brunstein, Lisa Ferro, and Beth Sundheim. ACE-2 Version 1.0 LDC2003T11. Philadelphia: Linguistic Data Consortium, 2003.

Anna Katarzyna Młynarczyk. *Aspectual pairing in Polish*. PhD thesis, Utrecht University, 2004.

Marc Moens and Mark Steedman. Temporal ontology and temporal reference. *Computational linguistics*, 14(2):15–28, 1988.

Alexander PD Mourelatos. Events, processes, and states. *Linguistics and philosophy*, 2(3): 415–434, 1978.

Alexander Nakhimovsky. Aspect, aspectual class, and the temporal structure of narrative. *Computational Linguistics*, 14(2):29–43, 1988.

Anna Nedoluzhko. Generic noun phrases and annotation of coreference and bridging relations in the Prague Dependency Treebank. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse (LAW-VII)*, pages 103–111, Sofia, Bulgaria, August 2013.

Alexis Palmer and Annemarie Friedrich. Genre distinctions and discourse modes: Text types differ in their situation type distributions. In *Proceedings of the Symposium on Frontiers and Connections between Argumentation Mining and Natural Language Processing*, Bertinoro, Italy, July 2014.

Alexis Palmer, Jonas Kuhn, and Carlota Smith. Utilization of multiple language resources for robust grammar-based tense and aspect classification. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, Lisbon, Portugal, May 2004.

Alexis Palmer, Elias Ponvert, Jason Baldridge, and Carlota Smith. A sequencing model for situation entity classification. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 896–903, Prague, Czech Republic, June 2007.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–106, March 2005. ISSN 0891-2017.

Rebecca J Passonneau. A computational model of the semantics of tense and aspect. *Computational Linguistics*, 14(2):44–60, 1988.

Rebecca J Passonneau and Bob Carpenter. The benefits of a model of annotation. *Transactions of the Association for Computational Linguistics (TACL)*, 2:311–326, 2014.

Philip L Peterson. On representing event reference. In *Fact Proposition Event*, pages 65–90. Springer, 1997.

Barbara Plank, Dirk Hovy, and Anders Søgaard. Learning part-of-speech taggers with inter-annotator agreement loss. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 742–751, Gothenburg, Sweden, April 2014.

Massimo Poesio. Discourse annotation and semantic annotation in the GNOME corpus. In *Proceedings of the 2004 ACL Workshop on Discourse Annotation*, pages 72–79, Barcelona, Spain, July 2004.

Livia Polanyi. The Linguistic Structure of Discourse. CSLI Technical Report, Stanford, CA, 1995.

Livia Polanyi, Chris Culy, Martin Van Den Berg, Gian Lorenzo Thione, and David Ahn. A rule based approach to discourse parsing. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, volume 4, Boston, Massachussetts, April 2004.

Rashmi Prasad, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind Joshi, Livio Robaldo, and Bonnie L Webber. *The Penn Discourse Treebank 2.0 Annotation Manual*, 2007.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K Joshi, and Bonnie L Webber. The Penn Discourse TreeBank 2.0. In *Proceedings of The 6th Language Resources and Evaluation Conference (LREC)*, Marrakech, Morocco, May 2008.

Sandeep Prasada. Acquiring generic knowledge. *Trends in cognitive sciences*, 4(2):66–72, 2000.

James Pustejovsky, José M Castano, Robert Ingria, Roser Sauri, Robert J Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R Radev. TimeML: Robust specification of event and temporal expressions in text. *New directions in question answering*, 3:28–34, 2003a.

James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, David Day, Lisa Ferro, Robert Gaizauskas, Marcia Lazo, Andrea Setzer, and Beth Sundheim. The TimeBank corpus. In *Corpus linguistics*, page 40, 2003b.

James Pustejovsky, Marc Verhagen, Roser Sauri, Jessica Littman, Robert Gaizauskas, Graham Katz, Inderjeet Mani, Robert Knippen, and Andrea Setzer. TimeBank 1.2, LDC2006T08. Web Download, April 2006.

Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning. A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 492–501, Cambridge, Massachusetts, October 2010.

Hans Reichenbach. *Elements of Symbolic Logic*, chapter The tenses of verbs. Macmillan, 1947.

Hans Reichenbach. *Elements of Symbolic Logic*. Dover Publications, 1980.

Nils Reiter and Anette Frank. Identifying generic noun phrases. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 40–49, Sweden, Uppsala, July 2010.

Josef Ruppenhofer and Ines Rehbein. Yes we can!? Annotating the senses of English modal verbs. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, pages 24–26. Citeseer, 2012.

Josef Ruppenhofer, Michael Ellsworth, Miriam RL Petruck, Christopher R Johnson, and Jan Scheffczyk. FrameNet II: Extended theory and practice, 2006.

Tanja Samardžíc and Maja Miličevíc. A framework for automatic acquisition of croatian and serbian verb aspect from corpora. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia, May 2016.

Beatrice Santorini. *Part-of-speech tagging guidelines for the Penn Treebank Project (3rd revision)*, 1990.

Sunita Sarawagi. Information extraction. *Foundations and Trends® in Databases*, 1(3): 261–377, November 2008.

Roser Sauri and James Pustejovsky. Factbank 1.0 ldc2009t23. Web Download. Philadelphia: Linguistic Data Consortium, 2009.

Roser Saurí, Robert Knippen, Marc Verhagen, and James Pustejovsky. Evita: a robust event recognizer for QA systems. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT-EMNLP)*, pages 700–707, Vancouver, B.C., Canada, October 2005a.

Roser Saurí, Jessica Littman, Bob Knippen, Robert Gaizauskas, Andrea Setzer, and James Pustejovsky. TimeML annotation guidelines, 2005b.

Sabine Schulte Im Walde. Experiments on the automatic induction of German semantic verb classes. *Computational Linguistics*, 32(2):159–194, 2006.

Diarmuid O Séaghdha and Simone Teufel. Unsupervised learning of rhetorical structure with un-topic models. In *Proceedings of the 25th International Conference on Computational Linguistics (Coling)*, pages 2–13, Dublin, Ireland, August 2014.

John R Searle. *Speech acts: An essay in the philosophy of language*, volume 626. Cambridge University Press, 1969.

Todd Shore and Alexis Palmer. Situation entity annotation manual. Unpublished manuscript, 2011.

Eric V. Siegel. Disambiguating verbs with the WordNet category of the direct object. In *Proceedings of Workshop on Usage of WordNet in Natural Language Processing Systems*, Universite de Montreal, 1998a.

Eric V Siegel and Kathleen R McKeown. Learning methods to combine linguistic indicators: Improving aspectual classification and revealing linguistic insights. *Computational Linguistics*, 26(4):595–628, 2000.

Eric Victor Siegel. *Linguistic Indicators for Language Understanding: Using machine learning methods to combine corpus-based indicators for aspectual classification of clauses*. PhD thesis, Department of Computer Science, Columbia University, 1998b.

Carlota Smith. The range of aspectual situation types: Derived categories and a bounding paradox. *Temporal reference, aspect, and actionality*, 1:105–124, 1995.

Carlota S Smith. *The parameter of aspect*, volume 43 of *Studies in Linguistics and Philosophy*. Springer Science & Business Media, 1997.

Carlota S Smith. Discourse modes: aspectual entities and tense interpretation. *Cahiers de grammaire*, 26:183–206, 2001.

Carlota S Smith. *Modes of discourse: The local structure of texts*. Cambridge University Press, 2003.

Carlota S Smith. Aspectual entities and tense in discourse. In *Aspectual inquiries*, pages 223–237. Springer, 2005.

Carlota S Smith and Mary S Erbaugh. Temporal interpretation in mandarin chinese. *Linguistics*, 43(4):713–756, 2005.

Radu Soricut and Daniel Marcu. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL-HLT)*, pages 149–156, Edmonton, Canada, May 2003.

Manfred Stede and Andreas Peldszus. The role of illocutionary status in the usage conditions of causal connectives and in coherence relations. *Journal of Pragmatics*, 44(2): 214–229, 2012.

Mark Steedman. *The syntactic process*, volume 24 of *Language, Speech, and Communication*. MIT Press, 2000.

William F Styler IV, Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C de Groen, Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova, and James Pustejovsky. Temporal Annotation in the Clinical Domain. *Transactions of the Association for Computational Linguistics (TACL)*, 2:143–154, 2014.

Sangweon Suh. Extracting Generic Statements for the Semantic Web. Master's thesis, University of Edinburgh, 2006.

Sangweon Suh, Harry Halpin, and Ewan Klein. Extracting common sense knowledge from wikipedia. In *Proceedings of the Workshop on Web Content Mining with Human Language Technologies at ISWC*, volume 6, Athens, Georgia, November 2006.

Simone Teufel. *Argumentative zoning: Information extraction from scientific text*. PhD thesis, University of Edinburgh, Edinburgh, UK, 1999.

Simone Teufel and Marc Moens. What's yours and what's mine: determining intellectual attribution in scientific text. In *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora*, pages 9–17, Hong Kong, October 2000.

Simone Teufel and Marc Moens. Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–445, 2002.

Simone Teufel, Jean Carletta, and Marc Moens. An annotation scheme for discourse-level argumentation in research articles. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics (EACL)*, pages 110–117, Bergen, Norway, June 1999.

Stefan Thater, Hagen Fürstenau, and Manfred Pinkal. Word meaning in context: A simple and effective vector model. In *The 5th International Joint Conference on Natural Language Processing (IJCNLP)*, pages 1134–1143, Chiang Mai, Thailand, November 2011.

Erik F Tjong, Kim Sang, and Hervé Déjean. Introduction to the CoNLL-2001 shared task: clause identification. In *Proceedings of the 2001 workshop on Computational Natural Language Learning (CoNLL)*, Manchester, UK, August 2001.

Milan Tofiloski, Julian Brooke, and Maite Taboada. A syntactic and lexical-based discourse segmenter. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP (ACL-IJCNLP)*, pages 77–80, Singapore, August 2009.

Kristina Toutanova and Christopher D Manning. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in NLP and Very Large Corpora (EMNLP/VLC-2000)*, pages 63–70, Hong Kong, October 2000.

Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL-HLT)*, pages 173–180, Edmonton, Canada, May 2003.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the Association for Computational Linguistics (ACL)*, pages 384–394, Uppsala, Sweden, July 2010.

Naushadand UzZaman, Hectorand Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. Semeval-2013 task 1: Tempeval-3: Evaluating time expressions, events, and temporal relations. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 1–9, Atlanta, Georgia, June 2013.

Benjamin D Van Durme. *Extracting implicit knowledge from text*. PhD thesis, University of Rochester, 2009.

Zeno Vendler. Verbs and times. *The philosophical review*, pages 143–160, 1957.

Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz, and James Pustejovsky. Semeval-2007 task 15: Tempeval temporal relation identification. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pages 75–80, Prague, Czech Republic, June 2007.

Marc Verhagen, Roser Saurí, Tommaso Caselli, and James Pustejovsky. Semeval-2010 task 13: Tempeval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval-2010)*, pages 57–62, Uppsala, Sweden, July 2010.

Henk J Verkuyl. *On the Compositional Nature of the Aspects.*, volume 15 D of *Foundations of Language, Supplementary Series*. D. Reidel Publishing Company, Dordrecht-Holland, 1972.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. ACE 2005 Multilingual Training Corpus LDC2006T06. Philadelphia: Linguistic Data Consortium, 2006.

Egon Werlich. *Typologie der Texte*. UTB für Wissenschaft, 1989.

Benjamin Lee Whorf. Grammatical categories. *Language*, 21(1):1–11, 1945.

Karina Wilkinson. Semantics of the common noun kind. In Gregory N. Carlson and Francis Jeffry Pelletier, editors, *The Generic Book*, Studies in Communication, Media, and Public Opinion, pages 383–397. University Of Chicago Press, 1995.

Jennifer Williams. Extracting Fine-grained Durations for Verbs from Twitter. In *Proceedings of the ACL 2012 Student Research Workshop*, pages 49–54, Jeju Island, Korea, July 2012.

Jennifer Williams and Graham Katz. Extracting and modeling durations for habits and events from twitter. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 223–227, Jeju Island, Korea, July 2012.

Naiwen Xue, Fei Xia, Fu-Dong Chiou, and Marta Palmer. The penn chinese treebank: Phrase structure annotation of a large corpus. *Natural language engineering*, 11(02): 207–238, 2005.

Nianwen Xue and Yuchen Zhang. Buy One Get One Free: Distant Annotation of Chinese Tense, Event Type and Modality. In *Proceedings of The International Conference on Language Resources and Evaluation (LREC)*, Reykjavik, Iceland, 2014.

Nianwen Xue, Hua Zhong, and Kai-Yun Chen. Annotating "tense" in a tenseless language. In *Proceedings of The International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco, 2008.

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Rashmi Prasad, Christopher Bryant, and Attapol T Rutherford. The CoNLL-2015 shared task on shallow discourse parsing. In *Proceedings of The SIGNLL Conference on Computational Natural Language Learning (CoNLL)*, page 2, Beijing, China, July 2015.

Robert Zangenfeind and Barbara Sonnenhauser. Russian verbal aspect and machine translation. In *Conference Proceedings of Computational Linguistics and Intellectual Technologies/Komp'juternaja lingvistika i intellektual'nye tehnologii*, volume 13, pages 700–709, 2014.

Alessandra Zarcone and Alessandro Lenci. Computational models for event type classification in context. In *Proceedings of The International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco, 2008.

Yuchen Zhang and Nianwen Xue. Automatic Inference of the Tense of Chinese Events using Implicit Linguistic Information. In *Proceedings of The Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, 2014.

Mengfei Zhou, Anette Frank, Annemarie Friedrich, and Alexis Palmer. Semantically Enriched Models for Modal Sense Classification. In *Proceedings of Linking Models of Lexical, Sentential and Discourse-level Semantics (LSDSEM)*, Lisbon, Portugal, September 2015.

Appendix

A Corpus sections: by annotators

corpus	**genre**	**annotators**
MASC	blog	A, C, D
MASC	email	A, B, D
MASC	essays	A, B, E
MASC	ficlets	A, C, D
MASC	fiction	B, C, D
MASC	govt' docs	A, B, D
MASC	jokes	A, B, D
MASC	journal	B, C, D
MASC	letters	A, B, C
MASC	news	B, C, D
MASC	technical	A, B, D
MASC	travel	B, D, E
Wikipedia	–	A, B, C