Automatic Error Detection in Non-native English

Rachele De Felice
St Catherine's College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Trinity 2008

Acknowledgements

While the responsibility of researching and writing up my DPhil thesis fell to me alone, there are many people in the background who have contributed in one way or another and have supported my efforts. This thesis would not have happened without them, and I would like to acknowledge their help and thank them here.

Several parts of this work have been presented in various fora, most notably the Flatlands meetings in 2006 and 2007, the 2007 ACL-SIGSEM Workshop on Prepositions, the 2008 CALICO Workshop on Automatic Analysis of Learner Language, and COLING 2008. The feedback and suggestions received on these occasions, and at seminars and talks, have been important in shaping the direction of my research.

Part of this work relied on the use of a subset of the Cambridge Learner Corpus, consisting of non-native English texts. This data has been made available to me by Cambridge University Press, whose assistance is gratefully acknowledged.

During the course of my graduate studies I have benefitted from the support of an Arts and Humanities Research Council grant. I am very grateful for their support, which allowed me to pursue my research unencumbered by financial worries.

On an individual level, many thanks are due to:

My supervisor, Professor Stephen Pulman, who first encouraged me to turn to computational linguistics and showed me how exciting the topic can be. Throughout my studies, he has been a tremendous source of help, leading me back on the right track with his insights and observations. I admire, and aspire to, his ability to rapidly identify the weak spots of an argument and offer suggestions as to where to look for improvements.

The NLP group and seminar attendees, in particular Stephen Clark and Pete Whitelock, for their illuminating comments and probing questions.

Laura Rimell for being wonderful and submitting herself without complaint to my random questions about numbers, LaTeX, judgements, results, and the meaning of life. I shudder to think what this past year would have been like without her around.

Joel Tetreault, who is always willing to discuss prepositions, learner errors, and the frustrations that come with them. It's nice to know there are others out there who think prepositions can be exciting.

Diane Nicholls for her kindness and availability in answering my numerous queries regarding the Cambridge Learner Corpus.

Kate Dobson and Robert Jubb, who have been invaluable in the early stages of my research in assisting with the error annotation of the learner data collected here.

Rita Calabrese, with whom I have shared ideas and conversations about L2 English, greatly broadening my perspective.

Nick Zair, who valiantly undertook to proofread the whole manuscript in an extremely short period of time. I am very grateful to him, and of course all remaining mistakes are my responsibility only.

Friends and family near and far (geographically) for being an incredible source of support. The Tuesday pub crew for giving me an excuse to relax (and compare word counts!) once a week. Especially my 'cheerleading squad', Nick, Sandra, and Samantha: thanks for hot meals, encouragement, ice cream, hugs and pats on the back, beer, words of support, chocolate, making sure I slept and took breaks, and generally keeping me sane. You have been amazing – thanks for being there. I'll try and reciprocate next year, but it's a hard act to follow...

My parents are a constant source of inspiration for me, for the enthusiasm and adroitness they display in always tackling new challenges and projects. I see this thesis, with its mix of science and humanities, as a synthesis of their interests, and it is to them that it is dedicated.

Abstract

This thesis describes the development of DAPPER ('Determiner And PrePosition Error Recogniser'), a system designed to automatically acquire models of occurrence for English prepositions and determiners to allow for the detection and correction of errors in their usage, especially in the writing of non-native speakers of the language. Prepositions and determiners are focused on because they are parts of speech whose usage is particularly challenging to acquire, both for students of the language and for natural language processing tools. The work presented in this thesis proposes to address this problem by developing a system which can acquire models of correct preposition and determiner occurrence, and can use this knowledge to identify divergences from these models as errors. The contexts of these parts of speech are represented by a sophisticated feature set, incorporating a variety of semantic and syntactic elements. DAPPER is found to perform well on preposition and determiner selection tasks in correct native English text. Results on each preposition and determiner are discussed in detail to understand the possible reasons for variations in performance, and whether these are due to problems with the structure of DAPPER or to deeper linguistic reasons. An in-depth analysis of all features used is also offered, quantifying the contribution of each feature individually. This can help establish if the decision to include complex semantic and syntactic features is justified in the context of this task. Finally, the performance of DAPPER on non-native English text is assessed. The system is found to be robust when applied to text which does not contain any preposition or determiner errors. On an error correction task, results are mixed: DAPPER shows promising results on preposition selection and determiner confusion (definite vs. indefinite) errors, but is less successful in detecting errors involving missing or extraneous determiners. Several characteristics of learner writing are described, to gain a clearer understanding of what problems arise when natural language processing tools are used with this kind of text. It is concluded that the construction of contextual models is a viable approach to the task of preposition and determiner selection, despite outstanding issues pertaining to the domain of non-native writing.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

> Choosing the right preposition is still hard for me [...] In addition to prepositions, I am often not sure about whether a noun is countable or not. And of course I will never completely understand the use of "the" and "a"!
> (Silva and Reichelt 2003:96)

The comment above, quoted from a Japanese student of English describing his experiences of learning and writing English, neatly encapsulates the problems this thesis sets out to address: namely, that prepositions and determiners are challenging parts of speech (POS) for both human and machine learners, and it is desirable to create systems capable of treating these POS and automatically detect and correct errors in their usage. This thesis describes the development of one such system, known as Dapper ('Determiner And PrePosition Error Recogniser'). This introduction explains the motivation behind the present research (Section 1.1) and sets out the objectives the thesis intends to fulfil (Section 1.2). In Section 1.3, the problems involved in the choice of a reference variety of language are briefly addressed; finally, Section 1.4 summarises the main contributions of this work.

1.1 Why prepositions and determiners?

The two POS which are the focus of this thesis are widely found to be particularly problematic for learners of English. Estimates of the frequency of errors involving them vary, but all reports place them among the most frequent error types. In a small corpus of non-native texts[1] collected and error annotated here as a preliminary study, for example, determiner errors represent 17% of the total and preposition

[1] Henceforth, non-native language will be referred to as L2, as opposed to native, or L1, language.

errors 12%. In Bitchener et al. (2005), these figures are 20% and 29% respectively. Therefore, in developing a system for automatic error detection for L2 writing, it seems desirable to focus on POS which are not only very problematic, but also, being function words, feature very frequently in even the simplest language productions.

Prepositions present such a challenge for learners because they can appear to have an idiosyncratic behaviour which does not follow predictable patterns, even across nearly identical contexts. For example, one says *look something up* ***in*** *a dictionary* but *look something up* ***on*** *the internet*, making it hard to appropriately generalise already acquired knowledge to novel instances. Words that are morphologically related also often do not occur with the same preposition: someone can be *independent* ***of*** *his parents*, for example, but *dependent* ***on*** *his wife*. Nor is this difficulty limited to similar contexts: even the same context can license more than one preposition, either with no change in meaning (*I got covered* ***in/with*** *mud*; *they met* ***at/in*** *primary school*), or leading to a different proposition (*there is a nice view* ***of/from*** *the mountain*; *he fell* ***by/in*** *the fountain*). This means that it is not sufficient to merely acquire individual lexical item-preposition pairings, especially for the development of natural language processing (NLP) tools for error correction: these need to be more sophisticated and aimed at general patterns rather than single lexical items, to allow for the inherent variation present in language.

Determiners pose a somewhat different problem. Firstly, there are wide-ranging cross-linguistic differences in the use of this POS, including the fact that some languages do not have determiners at all. In fact, these variations could be exploited to an extent to predict what form determiner errors will take in the writing of students of different L1s. For example, as many South-East Asian languages do not use determiners, one will find that the writing of these students largely underuses them. Conversely, Romance languages such as Italian and Spanish use determiners much more frequently than English, and so the writing of these students might be found to suffer from their over-use.

The other major problem in determiner occurrence is that, unlike prepositions, their choice is more dependent on the wider discourse context than on individual lexical items. The relation between a noun and a determiner is less strict than that between a verb or noun and a preposition, as the main factor in determiner choice – apart from whether a noun is countable – is the specific properties of the noun's context. For example, one can say *girls like football* or ***the*** *girls like football*, depending on whether one is making a general statement about all girls or just referring to a specific group of them. Similarly, both *she ate* ***an*** *apple* and *she ate* ***the*** *apple* are

grammatically well-formed sentences, but only one may be appropriate in a given context, depending on whether the apple has been mentioned previously.

These issues make it extremely difficult to define clear-cut rules describing every possible kind of determiner occurence. This is aggravated by the existence of several culture-specific rules regarding the use of definite determiners for nouns which are in some way part of shared knowledge, regardless of previous mention: for example, *the Queen* in the United Kingdom, understood at present to be Queen Elizabeth II, or *the Head of Department* within any one department, as there is a presupposition that this has a unique referent in each department. This kind of knowledge is very hard to encode, especially as it may vary greatly across contexts.

Knowledge of correct preposition and determiner usage is important not just in the domain of learner language, but in many NLP applications, too. Regarding prepositions, for example, Saint-Dizier notes:

> Prepositions turn out to be a useful category in a number of applications such as indexing and knowledge extraction since they convey basic meanings like instrument, means, comparisons, amounts, approximations, localizations, etc.
> (Saint-Dizier 2005:145)

Tasks involving natural language generation are also particularly affected by incorrect usage of either of these POS. For instance, in the output of machine translation, automatic summarization, or question-answering systems, one would aim to avoid generating incorrect sentences such as **I go at the university* or **the Italy is a beautiful country*. Indeed, some of the earlier work in this area was aimed at overcoming these issues in NLP applications working on L1 data rather than the error correction of human-produced text. The findings of the research presented in this thesis, then, can find a wider application beyond the domain of L2 writing[2].

1.2 Thesis aims

The work presented in this thesis aims to

1. develop a method for the automatic acquisition of usage models of prepositions and determiners in correct L1 English;

[2]This exchange of knowledge could also flow in a different direction: one could examine whether the issues encountered by human learners are similar to problems found in NLP tasks, and if so, whether any solutions proposed for learners could be in any way applied to these domains.

2. develop a system capable of detecting and correcting inappropriate preposition and determiner use, with a particular focus on L2 language;

3. highlight some issues related to the application of NLP tools to L2 language, and how these may affect performance and evaluation;

4. establish whether the use of contextual features including semantics and syntactic analysis has a positive impact on performance in these tasks;

5. derive insights into the occurrence patterns of prepositions and determiners which may be useful from linguistic and pedagogical perspectives.

The task set out in item 1, as discussed more fully in Chapter 3, may seem to run counter to the issues described in the previous section. However, the belief that "despite the very real complexity of prepositional collocation in English, there is vastly more system and logic here than is normally realised" (Lindstromberg 1995:17-18) is at the heart of this thesis. It intends to demonstrate the validity of this claim by identifying contextual features, for both prepositions and determiners, that can satisfy this task.

The L1 model developed will be at the core of the system mentioned in item 2. Despite the challenges posed by the use of NLP tools on learner language (see Section 2.1.2 and Section 6.1.2), using L1 models for the detection of anomalies in L2 language can be a successful strategy. It is inevitable that problems will arise, as noted in item 3; these will be discussed and possible solutions will be presented.

A key point for discussion in tasks involving the learning of contextual models for prepositions and determiners is the role played by the various features selected. As noted by Lee and Knutsson (2008), there is no detailed treatment in the literature of the contribution made by individual features. The models developed in the course of this research draw on several different features of varying semantic and syntactic complexity, including the use of full parsing, which does not usually feature in these tasks. As stated in item 4, an in-depth analysis of the value of each of these features will be provided, to understand what factors have the greatest importance, and whether the introduction of deep syntactic processing to the task is justified.

Finally, although the main objective of this thesis is the development of a successful NLP application, the resources collected in the course of the research may prove a rich source of insights into linguistic processes, as noted in item 5. This data can be used to inform studies on English as both a first and second language, for example highlighting areas where L2 writers are in need of particular attention.

What this research does *not* attempt is an evaluation of the usefulness of such an application for learners. Although it is possible that in the future DAPPER will be released as an interactive program from which students can receive immediate feedback and further clarifications, allowing for the tracking of their progress, during its development it has only been tested on L2 corpus resources, with no access to current students. The aim of this thesis is to assess the viability of developing such an application, not carry out a study on the validity of error feedback[3]. Furthermore, the issue of the merits or otherwise of error feedback are hotly debated: reviews of the main positions taken over the last decade can be found for example in Chandler (2003), Lee (2004a), Ferris (2004), and Bitchener et al. (2005), the latter finding that corrective feedback does not have a great impact on preposition errors. Regardless of this, a tool for error correction finds wide application not just in the L1 domain, as described in Section 1.1, but also for those who wish to verify the grammatical accuracy of L2 texts, be they students interested in improving their writing, or their instructors requiring a tool to assist them in their assessment.

1.3 Linguistic excursus: what English?

An important issue to address in this type of research is what variety of English is taken as 'correct', and what criteria are used to identify something as an error. Prepositions in particular are not immune to the dialectal differences found in English, especially between American and British English. While this is not necessarily a problem for the learners, as it is unlikely to affect the ease with which they can communicate, it can be a significant issue for this task. If the model of preposition usage acquired derives from British English, divergent uses which are acceptable or preferred in American English might be then marked as errors. When testing on learner data, this can cause false alarms, as one does not always know what variety of English the student has been taught. In the present research, it is attempted to minimise this risk by using data from the Cambridge Learner Corpus, which, comprising mostly exam scripts from the Cambridge Examinations Syndcate, is assumed to take British English as its reference. However, this highlights a general problem posed by working with learner text, namely the difficulty of controlling for all the possible variables.

[3]In taking this stance, the present work falls under the description given by Chapelle (2001:42): "Although AI and computational linguistics offer potentially useful software technologies, researchers in these areas are interested in developing computer programs rather than in developing learners' ability.".

Of course, this problem is not restricted to British vs. American English. Many different varieties of English exist, which may be standard in the student's home country or the country where most of their learning took place – Australian, Canadian, Indian, and so on. Although error annotators and corpus compilers can decide what variety of English to use as reference, it is not always possible to know what variety of English the learner has in mind when writing. What is an error in one variety may be acceptable in another: how can one tell what model the learner is aiming for? The issue of which English is 'right' affects not just NLP applications but the discipline of L2 English teaching overall. Different positions are advocated, generally aiming for "good quality L2 English, as spoken by academics and professionals" (Mauranen 2004:207)[4]. However, the position taken in this thesis is that it is important that, while retaining an appreciation of the value of geographical varieties, learners are given a clear description of what is considered 'correct' in mainstream varieties of English, especially where this information allows access to more prestigious jobs or educational opportunities.

From an NLP perspective, too, there is a lack of consensus on this matter, as described for example by Prat Zagrebelsky (2004:46). The author notes that, depending on the research question, there are at least four different ways to compare learner language to a target language norm: with a standard reference corpus such as the British National Corpus; with the production of 'expert' native speakers (e.g. newspaper editorials, academic texts); with equivalent native speakers' productions (for example essays by university students); with the language found in student textbooks. The last option may be the most fair, as this kind of L1 language is often the only kind learners have had exposure to. The view taken here is that by selecting an L1 resource with wide-ranging content, there is a higher likelihood of the model finding wider applicability beyond the L2 domain. However, the difference in domain and text-type between training and testing data may be found to impair performance, as will be discussed in Chapter 6.

1.4 Thesis structure and main contributions

With reference to the objectives of this thesis described in Section 1.2, the contributions of this thesis are as follows:

[4]See also James (1998:52ff.) for a discussion of the "unattainable ideal of the native speaker".

Acquisition of L1 models

The development of a machine-learning based approach for the automatic acquisition of usage models for nine preposition classes (***at, by, for, from, in, of, on, to, with***) and three determiner classes (***a/an, the, null***[5]) in correct, L1 English is described. These models are shown to achieve up to 70.06% accuracy on a preposition selection task in L1 data and up to 92.15% accuracy for an analogous determiner selection task.

Individual item analysis

Results from the L1 task for each individual preposition and determiner are analysed and discussed in depth, observing relations between particular contextual features and items and establishing whether there are inherent difficulties in the learning of any of the target items. Detailed studies of this kind are rare and this data can serve as the basis for a variety of linguistic, NLP, and pedagogical resources.

Application to error correction

The L1 usage models are applied to L2 data to assess performance on an error correction task in this domain. Dapper achieves average precision of 42% and average recall of 35% in recognising preposition selection errors. On determiner errors, accuracy is up to 67.5% in the case of errors involving confusion between ***a*** and ***the***. A detailed analysis of the results is also given, addressing the major issues arising from using NLP tools on L2 data, particularly for the task of error correction.

Feature analysis

An in-depth assessment of the role played by each feature is carried out, quantifying its contribution from statistical and linguistic perspectives. There are very few studies of such length in the field; the present one is particulary important as it makes use of full parsing, unlike other approaches which do not carry out deep syntactic processing.

Linguistic and pedagogical insights

The analysis of individual features and items offers the opportunity to reflect on the mechanisms underlying the use of prepositions and determiners in the language, confirming or amending previously held views on the matter. Comparisons between the

[5]Throughout this thesis, **null** is used to refer to the absence of determiner, as in *she baked [null] cakes yesterday.*

contexts of use of individual items in L1 and L2 data are also made, helping us understand what requires the attention of L2 instructors and of NLP application developers.

In carrying out this research, a wealth of data on preposition and determiner occurrence has been acquired in the form of contextual patterns and feature frequency. As a long term goal, it is envisaged that this data can be turned into a resource providing information for these POS on aspects such as co-occurrence frequencies, syntactic frames, lexical preferences, and so on.

The structure of this thesis is as follows: in Chapter 2, the present research is put in context by reviewing related work in the fields of L2 language generally and preposition and determiner selection more specifically. In Chapter 3, the methodology adopted in developing the L1 models is described: the motivation underlying the choice of feature set is explained, along with the data and tools used for their training. The results on L1 data are presented in Chapter 4, including variations on the training parameters and discussion of results on each individual preposition and determiner. Chapter 5 gives the results of the feature analysis, discussing the contribution of each feature individually, while the results and discussion of the L2 error correction tasks are found in Chapter 6.

Chapter 2

Background and related work

This chapter puts the present research into context by offering an overview of current and recent work in the field of L2 research, with a special emphasis on prepositions and determiners. It begins by introducing some of the general issues involved in working with L2 text (Section 2.1) and reviewing other approaches to error detection (Section 2.2). In Section 2.3, related work on prepositions (Section 2.3.2) and determiners (Section 2.3.1) is discussed, with regard to both L1 and L2 contexts.

Over the past twenty years, there has been a flourishing of studies centred around the application of NLP techniques to L2 data. The main areas of focus of these can be grouped under three categories. The first is the creation of 'learner corpora' collecting students' L2 English texts. These can be used to analyse several aspects of the learning process and help researchers understand the steps involved in the acquisition of a foreign language; they will be discussed more fully in Section 2.1. Related to this is the development of tools to assist students and teachers in the detection and correction of errors, such as the one presented in this thesis; this topic will be addressed in Section 2.2. There is also a wealth of research on the use of L1 corpora to enhance and support the students' learning experience by giving them access to examples of 'language in use', allowing them to discover patterns on their own, and developing their linguistic sensitivity through hands on experience (this is generally referred to as CALL, 'Computer Assisted Language Learning'). The research presented in this thesis does not touch upon this subject so it will not be discussed further; good overviews of the topic may be found in Chapelle (2001), Aston et al. (2004), and Chambers (2005), and most recently in papers in the 2007 *Annual Review of Applied Linguistics*, especially those by Blake (2007), Chapelle (2007), and Douglas and Hegelheimer (2007). Indeed, as explained in Chapter 1, in this thesis there is no debate on the pedagogical merits or otherwise of DAPPER, and whether it

can in fact assist in the learning of preposition and determiner usage. What is being presented is merely an example of a practical application of NLP tools and L1 and L2 corpora to suggest a possible approach to the issue of error detection.

2.1 Working with learner language: L2 corpora

Like L1 corpora, L2 corpora are digitalised[1] collections of language samples, usually consisting of written text, produced by non-native rather than native speakers. Having such computerised resources available presents many advantages: a large amount of data can be stored and analysed in greater depth and in a wider range of ways than could be possible by hand, or just by relying on the intuitions of instructors. A detailed overview of research in the field can be found in Granger (2004) and Granger et al. (2007), among others; in this discussion, the focus will be on those areas which are most relevant to the work presented here.

Naturally, language instructors and researchers have long been interested in looking at learner productions to inform the development of teaching approaches and theories about second language learning. However, the creation of L2 corpora as digital resources which can be annotated in several different ways allows for systematic and complex analyses. In the discussion that follows, some of these approaches to annotation will be examined.

One of the most important features of L2 corpora is the presence of a rich set of metadata characterising the author of each text and the context of its production. Several researchers have compiled lists of the type of information that should be included in these descriptions; those found in Granger (1998a:7-9) and Granger (2004) are representative of many of them. They include elements such as the student's L1, education level, years of instruction in English, age, country of origin, task setting (e.g. timed exam, take home assignment) and so on. Similar considerations are made by Ellis and Barkhuizen (2005:30), who state that the minimum information required in a description of the data should include the learners' social and situational background, the context in which the writing task was carried out, its genre and topic, the timing, and whether any reference tools were used. The availability of this metadata is a valuable resource for researchers, as it allows the investigation of complex feature interactions, the possibility to filter data according to a particular characteristic (e.g.

[1]Either because the students have typed their work, or because their handwritten essays have been transcribed into digital form – although this presents potential dangers as the transcribers have to be careful to preserve any errors originally present and not introduce any spurious ones themselves.

L1), or to focus only on one level of proficiency. In fact, it could be argued that without this corollary information, the corpus loses much of its value as it would be hard to know how generally applicable any conclusions drawn from its study might be. As Granger (2004:125) notes, "failure to control these factors [i.e. the learner and essay characteristics] greatly limits the reliability of findings in learner language research".

At the same time, however, this abundance of variables also creates several difficulties. Unless everyone works with the same corpus resources, for example, it is almost impossible to replicate the same set of conditions in different settings. Furthermore, if one wishes to follow relatively strict criteria in building a corpus, and create a resource that is not too diversified across the several parameters, it is hard for individual researchers to develop corpora of the scale of those available for L1 studies, especially for English. Indeed, as shown below, the very large L2 corpora which do exist are the product either of wide-ranging collaborations or the efforts of publishers and educational companies rather than smaller academic groups.

Finally, there exists another important difference between L1 and L2 corpora, namely their ultimate goal. In L1 studies, large-scale resources are usually advocated to allow the discovery of statistically significant patterns and generalisations. In L2 studies, however, given the multiple dimensions of variations that exist, small scale studies may be of assistance even if they do not lead to discoveries which can be held to be universally valid. For example, the results of such studies may suggest to instructors what aspects of language require more attention, or they may help inform the directions taken by the development of NLP tools, especially those focused on particular items or user groups. As Granger observes:

> A teacher analysing his learners' output with the help of computer techniques may well come up with highly interesting new insights based on quantitative information which may in itself not be statistically significant but which nevertheless has value within a pedagogical network. (Granger 1998b:15)

2.1.1 Main L2 corpora

Researchers often compile small scale corpora to fulfil their needs for a particular task or domain, and it would be impossible to review them all here. A thorough and comparative overview of several corpora is given in Pravec (2002), and a detailed

table summarising the characteristics of many of these is found in Prat Zagrebelsky (2004:51-53). In this section only the most well-established corpora are introduced.

The International Corpus of Learner English (ICLE)[2] is based at the Université Catholique de Louvain in Belgium and is one of the largest academic learner corpora. It is considered the first learner corpus created in an academic setting, its compilation having begun in 1990. The texts contained in it are by advanced university students and belong to one of the typical genres of L2 writing, namely essays arguing for/against a position relating to literary topics or current affairs. The students display a variety of L1 backgrounds, but the compilation of the corpus is strictly controlled (for example, the same list of possible essay topics is given to all the students). This means that there is a great deal of similarity within the learner characteristics, making it possible to carry out a variety of comparative studies and avoiding the pitfalls described above. The overall corpus is of considerable size (currently over 3 million words), but the individual components (one for each L1) are only of around 200,000 words each. While the attention to compilation criteria is comparable to that found for large L1 corpora such as the British National Corpus, it is clear that, in terms of size, L2 corpora tend to be on a much smaller scale.

Commercial learner corpora are very large-scale undertakings, comprising several million words, which are compiled with the aim of assisting in the preparation of material such as learners' dictionaries, textbooks, and other pedagogical resources. Two well known ones are the Cambridge Learner Corpus (CLC)[3] and the Longman Learners' Corpus[4]. The former is the source of the L2 data used in the present research, and is described in greater detail in Section 6.1. Both are characterised by the inclusion of exam scripts (and essays, in the case of the Longman Corpus) from all over the world, representing a wide range of L1s, topics, task types and proficiency levels. The size and breadth of these corpora makes them a rich source of information; however, because of commercial interests, they are not usually freely available to the academic community.

Other academic corpora which are gaining widespread use are the Chinese Learners of English Corpus (CLEC)[5] and the Hong Kong University of Science and Technology (HKUST)[6] learner corpus. The former contains around 1 million words, while

[2]http://www.fltr.ucl.ac.be/fltr/germ/etan/cecl/Cecl-Projects/Icle/icle.htm. Unless otherwise indicated, all URLs last accessed in September 2008.

[3]http://www.cambridge.org/elt/corpus/learner_corpus2.htm

[4]http://www.pearsonlongman.com/dictionaries/corpus/learners.html

[5]http://langbank.engl.polyu.edu.hk/corpus/clec.html

[6]No URL available.

the latter is much larger at over 25 million words. In both cases, they consist of texts by learners of one L1 only, which may be seen as a limitation or as a useful complement to other corpora which may be more slanted towards European L1s.

2.1.2 Using NLP tools on L2 corpora

Although a learner corpus in its 'raw' form is a useful resource, its real value lies in the kind of information that can be extracted from it. Tasks such as examining rhetorical features or compiling lists of word frequencies do not require particularly complex processing of the data[7], and can yield useful insights into the lexical sophistication and richness of vocabulary of the students, or issues such as the under- or over-use of phrases and words. This can be compared to that of native speakers of a similar age and education to assess progress. However, like any corpus, a learner corpus is more informative if we can easily extract information from it about POS, syntax, and other grammatical and semantic features, using any of the several NLP tools available. Unfortunately, the nature of learner corpora, and their divergences from native English ones, means that these tools do not always perform as well because they are not trained to deal with the non-standard language used in the corpora. A good introduction to these issues, which will be briefly addressed below, is found in Meunier (1998).

POS tagging, for example, often relies on statistically observed sequences of certain parts of speech as well as lexical item information; if the learner language is very different from correct English, presenting for example unexpected word order, it is unlikely the tagger can perform well on this input. These, and lemmatisers, also use heuristics based on the spelling of the word, so ill-formed or misspelled words are likely to cause confusion and, in the case of the lemmatisers, incorrect lemmas will be retrieved. The issue of POS tagging of learner text has been addressed by several participants in the ICLE project, as one of the aims of the project is to have all its text POS tagged using the TOSCA-ICLE tagger (de Haan and van Halteren 1997). Early results on the Dutch section of the ICLE, for example, reported a tagging success rate of 95% (de Haan 2000) despite encountering issues related to the presence of misspelled or incorrectly used words (e.g. the adjective *proud* instead of the noun *pride*), suggesting that performance is broadly good and tagging can add useful information to the data. On the other hand, van Rooy and Schäfer (2003) discuss the results of using the tagger to parse data from the Tswana (a South

[7]Although the presence of misspelled or wrongly inflected words may lead to anomalies and omissions in the latter task.

African language) component of the ICLE and find it only achieves accuracy of 88%. However, this might partly be due to the fact that the TOSCA tagset comprises over 200 tags, so there is a greater potential for similar tags to be used incorrectly. As a general conclusion, it can be claimed that the benefits derived from having POS tagged learner data outweigh the problems encountered.

A closely related issue to POS tagging is that of parser performance. Since parsers usually rely on POS tag sequences, and are trained on, and expect, well-formed input, there is a high probability of incorrect parses being produced if the L2 input diverges greatly from typical L1 data and/or incorrect POS tags have been assigned. This is a real possibility: sometimes learner input is so ill-formed that even human readers struggle to understand its sense and structure, so it not surprising that NLP tools should find them challenging, too, such as in the following example taken from the Cambridge Learner Corpus:

(1) One on the most problems that Television have been created since that this had invested, it's the reduction.

The issue of parsing L2 text is addressed by de Mönnink (2000), among others. While clearly the addition of syntactic annotation allows the analysis of a wider range of phenomena in learner writing, such as the over-use or under-use of particular syntactic structures, the paper's author also acknowledges the presence of several obstacles. As well as impossible-to-parse sentences, as noted above, these include mismatches in number between subjects and verbs, and spelling and lexical errors which mislead the tagger and therefore the parser. In light of these issues, many researchers working with L2 data choose not to fully parse text, as seen in the discussion of work on error detection, because of the possibility of the parser output being unreliable. The approach proposed here, on the contrary, does make use of parsed learner data; as shown in Chapter 6, it is believed this is possible without impairing performance.

2.1.3 Error annotation

In addition to the markup shared with L1 corpora, the distinguishing and most attractive characteristic of learner corpora is the fact that they can be used to learn what kinds of errors L2 writers are most prone to, taken either as a general population or only considering subsets by L1 or proficiency level, for example. This systematisation of the account of learner difficulties is useful both for those developing learning materials and for individual instructors who can tailor their teaching to the specific problems of a particular L1, proficiency level, and so on (and conversely, knowing

which areas pose fewer problems allows one to know which areas need *less* focusing on – it is important to remember that these corpora are not just a collection of errors, but also a marker of progress). Indeed, it is often remarked that an L2 corpus without error annotation is of little or no value.

The need for the classification of error types had of course been recognised long before the availability of computerised resources[8], but the possibilities opened up by having computerised corpora which can be searched and cross-referenced have allowed this to be undertaken on a much larger scale.

For this kind of analysis, an extra layer of annotation is needed, namely error tagging. This process is far from straightforward. Firstly, the nature of the error itself has to be taken into account. Although this varies depending on the student's proficiency, often when some part of the text is erroneous the cause lies in more than one element. One must therefore decide whether to highlight one error over the others, or to capture all of them in some way. This of course depends on the choice of error taxonomy used. Several have been proposed, all of them sharing the basic principle that both the POS involved in the error and the type of error (e.g. omission, unnecessary items, incorrect selection) have to be recorded in some way. A good overview of the general criteria involved in such taxonomies, without reference to a specific scheme, is found in James (1998:102-114), where the various dimensions involved in characterising an error are described.

As reasons of space make it unfeasible to describe all the different error tagsets in existence, the discussion here will focus only on the two best-known ones used for English, which accompany two of the large-scale projects described in Section 2.1.1, the ICLE and the CLC. Both have very rich and comprehensive tagsets, which use combinations of letters to identify different types of errors and the parts of speech involved. The tags cover all the necessary categories: form, grammar, lexis, lexico-grammatical errors, style, register, word order, and so on. As both corpora include texts from many different L1 learners, the tagsets are not biased towards any particular L1.

The ICLE tagset (Granger et al. 2002) comprises 52 tags based around 7 major categories: form, grammar (general rules of grammar related to POS), lexico-grammar (POS complementation, use of dependent prepositions, count/uncountable nouns), lexis (semantic and collocational properties, subcategorisation), word (repetitions or omissions of words, wrong word order), register, style (longer stretches of

[8]For example in the work of Corder (1981) who identifies categories of learner errors to which we still refer today: omission, addition, selection, and ordering, although he does not find them sufficiently "deep or systematic"(Corder 1981:37).

text, incomplete or unclear passages), punctuation (whether missing, redundant or misused). Each category code can then be enriched with information about the POS of the particular error.

The CLC tagset follows the Cambridge Error Coding System devised devised and applied manually to the data by Cambridge University Press. Details of the scheme can be found in Nicholls (2003); in summary, for errors involving single lexical items the coding scheme records both the POS of the item, using a letter as mnemonic, and the type of error ('M' for missing, 'U' for unnecessary, 'R' for wrong selection); so, for example, 'MV' denotes a missing verb error, 'UT' an unnecessary preposition, and so on. There are also further codes for other kinds of errors involving more than one word such as 'CE' (complex error, where the intended sense cannot be established), 'CL' and 'ID' for errors involving collocations and idioms respectively, 'W' for errors relating to word order, and 'X' for cases where a negative has been incorrectly formed. An example sentence with its annotation is given here:

(2) **learner sentence:** It was properly work two three days, then it damaged other side of pipe.
with error annotations: It #UV was /#UV #W properly #TV work_worked /#TV _worked properly /#W #MT _for /#MT two #MT _to /#MT three days, then it damaged #MD _the /#MD other side of #MD _the /#MD pipe.
corrected version: It worked properly for two to three days, then it damaged the other side of the pipe.

The two annotation schemes are very similar and differ mainly in the nomenclature used for the various errors. The main drawback of these schemes, and indeed most error annotation projects, is that tagging is carried out manually, so the development of large-scale resources is forced to proceed slowly to account for training time and the physical limitations imposed by having human annotators. Furthermore, in many cases, especially where lexical errors are concerned, one is subject to the annotators' personal preferences in deciding what is right and wrong. As shown in Chapter 6, this is not a trivial issue; Tetreault and Chodorow (2008a) discuss how this problem can affect evaluation of error detection tasks. Wible et al. (2003) suggest a solution in the form of bootstrapping error annotations from a set of errors manually noted by teachers, to identify further errors of the kind a particular teacher or researcher is interested in, but the results are not conclusive and are tested only on two kinds of errors. For the time being, manual error annotation remains the main source of this data; however, it is useful to have some information available about the annotators,

for example what variety of English they are using as reference, to help interpret unclear or controversial annotation decisions.

2.2 Error detection

This section gives an overview of work on using NLP techniques for error detection, before focusing more closely on prepositions and determiners in the next section.

There is a large body of work on error detection, covering errors both in L1 and L2 writing. These studies range widely in scope from small projects on specific elements to broader research efforts covering several different aspects of the language. A full account cannot be given here for reasons of space; good overviews are found in Dodigovic (2005:especially Chapter4, pp. 85-139) and in Thomas (2004).

There is much debate on the issue of whether automatic error detection is at all viable. Granger et al. (2007) present an in-depth discussion of the topic, reviewing the main positions found in the discipline. Many are not very optimistic, and lament the lack of coverage or accuracy of the tools currently available. For example, the authors give the following quote from Amaral and Meurers (2006): "processing completely free production input, allowing any number and type of errors, is not tractable". This view justifies the approach chosen for the present work to focus only on errors pertaining to two particular POS although, as shown in Chapter 6, it is often very difficult to isolate individual errors and deal with them successfully. Indeed, all the work described below follows the approach of only focusing on one or two error types, too. It is important to highlight that the focus here is only on errors in written language; the correction of spoken language output, such as that of intelligent tutoring or automatic speech recognition applications, is not addressed.

Early, and sporadic, studies centred mostly on detecting 'real-word' errors, grammatical or spelling errors resulting in other legitimate words of the language (e.g. *desert - dessert, their - there*). One such early effort is described in Atwell (1987), where grammatical errors are flagged in English text on the basis of infrequent or unlikely POS tag sequences – particularly useful in the case of spelling mistakes resulting in other legitimate words. An unlikely sequence would indicate the presence of an extraneous, and therefore erroneous, element. This notion is not tested extensively, partly because of processing limitations at the time. It is, however, important to note that it was already recognised at the time that error recognition by means of ad-hoc 'mal-rules' (individual rules formulated to describe errors) may not be the most effective, and that the identification of divergences from a correct usage model

is a reliable indicator of the presence of an error or anomaly: this is also the approach taken by the work in this thesis.

Golding (1995), too, approaches the problem of discriminating among often-confused words, which he proposes to solve by looking at the surrounding lexical items, surrounding syntactic patterns, and by using these as features to train Bayesian classifiers to select the correct target word. Central to this task is the choice of appropriate contextual features, and establishing what dimensions of the context are most informative for this type of task. These are concerns shared by the present work, as choosing a target word from a set of preposition or determiner choices can be thought of as a comparable task. More than one aspect of the context is included, which is in agreement with the decision made here to consider lexical, syntactic, and semantic features (cf. Chapter 3).

The real-word error issue is also tackled using transformation-based learning, as described for example in Mangu and Brill (1997). Potentially confusable words are grouped in sets, from which the appropriate one is selected on the basis of surrounding lexical items and POS tags. The authors test their approach on a small number of confusables and achieve a success rate of up to 92%. While the small scale of their task limits any claims of more general applicability, the intuition to use several surrounding lexical items is an interesting one as it indirectly gives an idea of the topic of the text, which in turn can affect the choice not just of lexical items, but also of function words. Lack of world knowledge is a problem encountered in the present research, especially with regard to determiners: their approach, while not explicitly encoding such knowledge, may go some way towards addressing the gaps created by its absence, by selecting one domain rather than another, for example.

Chodorow and Leacock (2000) present an approach to the automatic identification of inappropriate use of particular lexical items in student essays. The method proposed by the authors is in many ways comparable to that used in the present research, despite focusing on a different type of error. It relies on local contextual cues around the target word to identify errors, where errors consist of divergences between the contextual models acquired from correct L1 text and the actual text observed – more specifically, use of unnecessary or missing function words in the immediate proximity of the target word, incorrect quantifier, and/or incorrect pluralisation. Function words and POS tag sequences make up the feature set. Their method is tested on 20 words such as *affect*, *aspect*, *concentrate* and *culture* and identifies errors around 78% of the time. Among the issues they identify as affecting performance are a lack of a semantic dimension, domain mismatch, incorrect POS tagging, and lack of parsing.

As will be seen in the discussion of DAPPER's results (Chapter 6), some of these problems are recurrent in using NLP tools on learner data. Others, such as the lack of a richer syntactic or semantic analysis, are issues that the present work tries to address.

It is clear from the discussion above that function words, such as prepositions and determiners, have not received as much attention in the past; nor have other types of grammatical errors, until more recently. In Foster and Vogel (2004) and Foster (2004), for example, a parser is used as the main tool in the detection of syntactic errors in L1 English. An important difference between this and much of the other work cited in this chapter is its use of error rules. A set of sentences containing syntactic errors is collected, and a parallel set consisting of the corrected versions of those sentences is created at the same time. The operation(s) used to correct the sentence are also logged, to gain a picture of the kinds of corrections required most frequently. These can then be referred to in the construction of error rules. The error rules are used to enable parsing in that, if a conventional grammar fails to parse the sentence, the error rules are invoked to establish whether with the application of any of them a parse can be found instead. Accuracy is around 84%; the main problems identified centre around the fact that sentences may have more than one error, while the recovery algorithm only contemplates the possibility of one error per sentence. While this is an interesting approach for parsing ill-formed native-like data, its limitation regarding the number of errors per sentence makes it less applicable to L2 data, which tends to contain several concomitant errors.

Similar issues arise in Wagner et al. (2007), where the task is to identify ungrammatical sentences using both POS tag sequences and more sophisticated linguistic processing. The two approaches are tested using a corpus of artificially created errors, to overcome issues of data sparseness. These are of four types: missing word, extra word, real-word spelling error, and agreement errors. The tools used include ranking the likelihood of POS tag sequences and running the sentences through a parser based on the Lexical Functional Grammar formalism, which can distinguish and rank optimal and suboptimal parses. These methods are also used in combination, yielding an average accuracy of 66%. As in the work above, this kind of artificially created error data is unlikely to present the same level of complexity found in real learner text, so it is unclear how well the results would translate to L2 data. On the other hand, it is encouraging to see work using more sophisticated NLP tools such as parsers, as it is important to assess the role they can play in this type of text.

Along these lines is also the work described in Andersen (2007), which investigates the possibility of using a naive Bayes classifier to identify whether L2 sentences do or do not contain any errors. Each sentence is represented by a set of features, ranging from POS tags and lexical items to grammatical relations (GRs): the learner data is run through the RASP parser (Briscoe et al. 2006) to extract this more sophisticated information. Using this approach, 70% accuracy is achieved; among the various kinds of errors, spelling mistakes and derivational errors are found to be the ones most easily identified. The paper also focuses more closely on the identification of two particular errors, namely the confusion between *a* and *an*, and determiner-noun agreement errors involving *this* and *these* (e.g. **this are my friends*). This paper is a further example of how parser output can be integrated in error detection at varying levels of complexity, especially in its use of GRs, which are an element included in the feature set of the present work, too.

Work by Brockett et al. (2006) takes a different approach to another kind of grammatical error, mass noun errors found in the Chinese Learner Error Corpus. Their work starts from the observation that errors often do not occur in isolation, but are usually bound up with other grammatical, lexical, or stylistic errors. They therefore propose to view error correction of L2 English as a special kind of machine translation and frame it in terms of a phrasal Statistical Machine Translation problem: the identification of an error is triggered by any difference between the user input and the model target sentences stored by the system. This work focuses on 14 frequently misused mass nouns; in testing, the system was able to correct 61.81% of mass noun errors found in data collected from the web. The problems arising from the presence of interrelated or concurrent errors, which form the motivation for the authors' research, are also encountered in the present work, as will be discussed in some detail in Chapter 6. This is particularly evident in the determiner task, where it is observed that errors in noun number and agreement are tightly bound with determiner errors in L2 writing, resulting in noun phrases (NPs) such as *he has a good interpersonal skill*. In errors like these, it is not always obvious that focusing only on the determiner element of the phrase is the best approach, as the determiner choice is dictated by the other, erroneous items in the NP. In this regard, the authors' suggestion of considering phrasal chunks rather than individual items may be more successful. The authors also advocate the need for more 'before and after' learner data, i.e. with and without corrections, as this is what is used to acquire mappings from erroneous phrases to corrected versions. In part this is already available with resources such as the CLC

and the ICLE which are error-annotated; on the other hand, it is not clear that this data is necessary if approaches based on L1 models only prove to be successful.

Often, these efforts on individual types of errors are pooled together to develop more complex applications which aim to provide a global assessment of student essays. This is the case for example of *Criterion* (Burstein et al. 2003; Attali and Burstein 2006) developed by Educational Testing Services[9], whose two components *e-rater* and *Critique* address grammatical, stylistic, and coherence issues within essays written in English by both L1 and L2 speakers. Microsoft Research has also recently released an experimental web application drawing on its research on error correction, the ESL Assistant[10].

Finally, it is important to remember that although most of the research in this area is focused on L2 English, doubtless because of the dominance of the language in the economic and cultural sphere, several other languages are also the target of error correction systems. For example, much work has been carried out on Swedish. In Hashemi et al. (2003), the task described is to identify grammatical errors in verb forms and agreement in the writing of L1 children; this is done by using difficulties encountered by a parser to flag potential error sites, similarly to some of the work described above. Bigert's work (e.g. Bigert 2004) focuses on real-word spelling errors using contextual information from POS tag sequences, simplifying them where necessary by removing adjectives and other modification, to allow underlying similarities to emerge. This approach is very similar to others seen above also focusing on this task; it is also similar in spirit to the one found in this thesis, in its intention to uncover more general underlying patterns of occurrence. Sjöbergh (2005) uses a reference corpus of chunked n-grams to detect errors in both L1 and L2 text, where errors consist of unseen sequences – however, this does not perform very well because of significant differences in text type between training and testing, which is an issue requiring careful consideration in developing a system with wide applicability. Further work on Swedish, in the context of the development of a grammar checker known as Granska, can be found on the project's webpage[11].

A different type of task is that undertaken by the developers of the ICICLE project[12], whose target users are native American Sign Language speakers writing in English. As their writing contains many errors of a different kind from those usually made by native speakers of English, this task is considered analogous to other

[9]http:/www.ets.org/research/erater.html
[10]http://www.eslassistant.com
[11]http://www.csc.kth.se/tcs/projects/granska/index-en.html
[12]http://www.eecis.udel.edu/research/icicle/

examples of English L2 error correction tasks. Underpinning this system is a set of 'mal-rules' corresponding to the kinds of syntactic errors learners at different stages of proficiency are most likely to commit, thus representing a departure from much of the other work presented in this section, which tends to identify errors only on the basis of divergence from a known correct L1 usage model.

2.3 Preposition and determiner errors

As is clear from the discussion in the previous section, error detection has been only recently shifting its attention to errors other than those involving incorrect lexical items, with the last year in particular (2008) seeing a flurry of activity on preposition selection. This section introduces a number of studies on the development of systems for the correction of preposition and determiner errors. Most of this research follows an approach similar to the one proposed here: a model of correct usage (which can be referred to as preposition/determiner selection or generation) is developed from L1 data, and is then applied to L2 data for error correction, where errors are identified on the basis of divergences from the correct usage models. Several aspects of these studies, such as feature selection, methodology, and results on L1 and L2 data, are best addressed in direct comparisons with the relevant aspects of the present work, so they will not be described in detail in this section. More complete discussions of their respective feature sets are found in Section 3.1.4; L1 results are compared in Sections 4.1.1 and 4.1.2 for prepositions and determiners respectively, and L2 results in Sections 6.3 (prepositions) and 6.4 (determiners). An important consideration to be made here is that there are severe limitations to the extent to which these studies can be compared to each other, and to the work in this thesis, as different data sets are used in each case for both training and testing.

2.3.1 Determiners

There are some early attempts at determiner generation in L1 English, for example in Knight and Chander (1994). The motivation for this work is the improvement of the output of Japanese-English machine translation rather than of L2 writing, but the core task – knowing which contexts require which determiner – is the same. It also provides a good example of how the components developed for these systems can be applied to both L1 and L2 domains, as suggested in Chapter 1. A major difference between this paper and most of the other work presented in this section (as well as this thesis) is that the null case is not considered, so the classification task is a binary

one, a choice between definite and indefinite determiner. The authors prefer the selection of contextual features rather than the creation of occurrence rules by hand, acknowledging the impossibility of handcoding real world knowledge and all possible exceptions into a system: "Leave the rules behind and move to a purely data-driven approach" (Knight and Chander 1994:782). This is the same view which underpins the present research. A decision tree is used for the classification task, yielding an overall accuracy of 78% on Wall Street Journal (WSJ) text. An interesting issue raised in the authors' discussion is the difficulty of measuring accuracy for determiner generation tasks. They note that success on the part of their system is deemed to be an instance of the system ouputting a choice which matches the original text. Hence, grammatical but unobserved decisions – which can be frequent, given that many noun phrases (NPs) can occur with either a definite or indefinite determiner – are scored as mistakes, giving the impression that the models acquired are not as thorough as they are. This is an issue also encountered in the analysis of the results presented here, as discussed in more detail in Section 6.3.2 and 6.4.2.

Minnen et al. (2000) is another early effort at determiner generation in consideration of its usefulness in domains such as machine translation, automatic summarization, and text-to-speech applications for people with disabilities. All three determiner classes are considered; their occurrence is learned using a memory-based learner on WSJ data. A range of lexical, syntactic, and semantic features is used, which will be described in more detail in Section 3.1.4.2; overall accuracy is around 83%. Among the salient aspects of this paper is the inclusion of semantic and deeper syntactic features in its contextual models, which distinguishes it from some of the other work addressed in this section in recognising that low-level information only may not be sufficient for good performance on this task. The models used in the present research also include these kinds of features, drawing however on different resources for their implementation. The paper also presents an analysis of the contribution made by each feature, as done here in Chapter 5; similarly to the conclusions reached in that chapter, it is found that several of the 'minor' features do not play as great a role as previously supposed.

A more recent discussion of determiner generation in a non-erroneous context is in Turner and Charniak (2007). This work, which is not explicitly aimed at error-detection applications, uses a language model for the task to select the most probable of the three possible candidates in a particular context. The model is trained on data from the Penn Treebank and from the North American News Text Corpus. The highest accuracy achieved is 86.74%; in its analysis of the mistakes made by the

system, the issues that arise are those shared by most other work on this topic: the fact that often more than one option is plausible, and the lack of world knowledge, which can override grammatical considerations. The results in this paper are considered state-of-the-art for the L1 generation task. In Section 4.1.2, they are discussed further with reference to the figures obtained by other studies.

Lee (2004b) introduces the determiner component of the system described in Lee and Seneff (2006), which is discussed in greater detail below (Section 2.3.2). The focus of this task is determiner generation; the approach used involves a variety of semantic and syntactic features (described in Section 3.1.4). It also stands out from much other work on the topic for its attempt at capturing discourse features, by tracking whether a given NP occurs in the previous five sentences. A maximum entropy classifier is trained on Penn Treebank data, which is also used for testing: errors are artificially introduced without affecting the remainder of the sentence. This means that the test data is rather different from typical L2 text, as the latter tends to display a variety of interrelated errors. This work is therefore more similar to papers on error detection in L1 output than on research on learner data. On error-free instances, accuracy achieved approaches 88%. Notably, the inclusion of a discourse feature, at least in this particular implementation, is not found to make the contribution that had been expected; this is a finding shared by Han et al. (2006), as discussed below.

Work on determiner error detection has also been carried out by Nagata and colleagues (Nagata et al. 2005a,b, 2006b), with a focus on L1 Japanese learners of English. Recognising the impracticality of manually creating error detection rules, which would never achieve comprehensive coverage, the authors propose instead, like other work in this section, the automatic derivation of rules from statistical contextual regularities. This takes two forms. The more simple approach requires only acquisition of determiner-noun probabilities on the basis of observed frequency (Nagata et al. 2005b), but it is found to be unsatisfactory for nouns which have very low frequencies or can be both mass and count nouns (e.g. *paper* - *read the paper* vs. *recycled paper is often brown*). This thesis shares the view proposed in Nagata et al. (2005a), namely that a more generally applicable system must focus on contextual patterns rather than simple lexical item co-occurrence statistics. In that paper, syntactic information is introduced, consisting of the heads of any verb phrases (VPs) or prepositional phrases (PPs) dominating the NP, as well as more abstract patterns using just the POS tag rather than the individual lexical items, to avoid data sparseness issues. While this is similar to the approach taken in the present work, it is also important to focus on the phrases of which the target NP might be a head, for example

if it is modified by any PPs (as discussed in Section 3.1.3.2), leading to the inclusion of a richer representation of the determiners' syntactic context in the feature set. On a small evaluation set of 250 errors, the model in Nagata et al. (2005a) achieves 64% recall and 77% precision, although it is not clear what kind of errors these are (replacement, omission, or unnecessary); spelling mistakes are corrected previous to the task. The authors find that the introduction of syntactic information makes a positive contribution to performance; this study will address the same issue but with a larger set and variety of test data, and these findings are expected to be confirmed.

Han et al. (2006) is a thorough investigation into the possibility of using a maximum entropy classifier to acquire models of determiner use and apply this information to error detection in L2 writing, making it most similar to the work presented in this thesis. The main conclusion is that this system performs at around 83% accuracy on correct L1 text (similarly to Minnen et al. 2000), while on the error detection task it achieves 52% precision and 80% recall. These figures are compared to DAPPER's results in Section 6.4. The authors recognise the need for a variety of factors to be considered, such as lexical properties of the noun, syntactic context, discourse factors, and even general knowledge, although not all of these are included in the feature set used by them: there is no semantic information, for example. The data is POS tagged and NP chunked but not parsed.

For evaluation on L2 data, essays written by three groups of L1 speakers are selected: Chinese, Japanese, and Russian, as these groups are found to encounter problems in determiner use with particular frequency. Restricting evaluation to certain subsets of L1s only may make results less directly comparable to the ones presented in this thesis, as all L1s available in the L2 data are included, and the distribution of error types may differ significantly among them. However, it is found that the relative infrequency of the confusion between ***a*** and ***the*** compared to the greater prominence of unnecessary and missing determiner errors observed in the data is also reflected in Han et al. (2006), suggesting that L1 differences may not have as dominant an effect as thought. There are further differences between this article and the present research which may affect evaluation. For example, NPs with misspelled words are not included in their test data. As shown in Chapter 6, such errors can be a significant hindrance to the classifier's good performance, so it is expected that their removal will improve results. Furthermore, in many cases of missing determiner errors the classifier's task is made easier because the human annotators have not specified which of ***a*** or ***the*** is required, giving it a higher chance of success – something not considered in

the present approach. Notably, the introduction of a discourse dimension is experimented with, to improve performance on errors due to confusion between ***a*** and ***the***. This consists of a simple method of tracking previous mentions of a noun in the text; it is however found not to bring any significant improvement to system performance. Perhaps, as the authors surmise, this is because the discourse dimension is composed of several factors other than previous mention (e.g. common knowledge) which may require a more complex representation. This finding supports the decision not to use discourse features in this first implementation of the system presented here.

Finally, Yi et al. (2008) adopt a different approach altogether to the issue of determiner error correction, by relying on web frequency counts only, having observed that L2 English speakers often use web searches as an informal way of addressing their language doubts (for example searching for two possible forms of what they intend to write and adopting the more frequent one as the correct one). In this system, each sentence is tagged and chunked to identify potential *loci* for determiners, which they term "check points". Then, for each check point, web queries are formulated using chunks or individual words as elements in the query, and each query is created in three forms, one for each possible determiner class. On the basis of frequency and usage ratios returned, the system establishes which combination is most plausible. On L2 data, it achieves 62.5% precision and 40.7% recall. Performance on collocation errors is found to be less good, leading the authors to conclude that "a web-based approach should be combined with local lingustic resources to achieve both effectiveness and efficiency". While the role web counts can play in many areas of NLP is recognised, the authors' observation that some linguistic input is required in tasks dealing with error correction is a key premise of this work. In the course of this thesis, it is proposed to show the role such linguistic input can have.

2.3.2 Prepositions

Work on preposition generation and error detection has been developing more recently. The former task is the topic of Gustavii (2005), whose aim is to improve the generation of prepositions in machine translation output. Having machine translation output rather than L2 writing as the target language is an important difference, as the former, in the context of this work, is estimated to be already correct on average 80% of the time, meaning that the quality of data is generally already high. However, there are also key similarities in the choice of contextual elements used to determine the appropriate selectional restrictions: the lemmas and POS of the surrounding lexical items are considered the triggers for preposition selection, which are features used

both in the present work and in other work discussed in this section. No syntactic analysis is performed to retrieve this information, which the author recognises can be a drawback, since the head and object of the preposition can only be approximated rather than determined with certainty:

> With fully parsed data, the governor, as well as the governed nouns, would be recognized with higher precision. The resulting classifiers would however be dependent on having access to fully parsed data, something which is not always output from rule-based machine translation systems.
> (Gustavii 2005:115)

This observation on the importance of more rigorous syntactic information supports the decision to include a full parse of the data in the present work.

In work by Chodorow and Tetreault (Chodorow et al. 2007; Tetreault and Chodorow 2008a,b), a system for automatic preposition error correction (both incorrect selection and unnecessary preposition errors) is described. This is based on a maximum entropy classifier which has been trained to recognise correct usage for 34 prepositions, using a set of contextual features which includes POS and lexical items. The work discussed in these papers shares many features with the present research, and more detailed comparisons of the relevant aspects will be given in Section 3.1.4.1 (feature set), Section 4.1.1 (L1 results) and Section 6.3 (L2 results). The training data here consists of texts from the MetaMetrics Lexile corpus, which contains materials aimed at American high school students. On L1 data, accuracy of 79% is achieved. Several filters to minimise the risk of false alarms – identifying the presence of an error where there is in fact none – are also introduced; these address issues which have been encountered in the present work, too, although as yet no provision has been made for them. They include skipping misspelled words, and special handling of cases where the preposition used in the text is either a possible antonym of the one given as correct (e.g. *to* vs. *from*) or is given a probability score very similar to that selected by the system, to account for cases where more than one preposition is acceptable in a given context. This system achieves up to 82% precision and 14% recall on incorrect preposition selection errors; when a component accounting for extraneous use is included, these figures rise to 84% and 19% respectively. In their work, the authors also address the issue of evaluation, such as the potential pitfalls of often having to rely on error annotation carried out by just one rater. This is a problem which has also surfaced in the present research, as will be discussed more fully in Chapter 6.

A similar perspective on preposition error detection is found in Lee and Knutsson (2008). Before attempting error detection itself, the authors address the issue of the acquisition of L1 models, similarly to other work discussed in this section. Here, however, the focus is particularly on the contribution made by various features, something which is addressed in detail in this thesis in Chapter 5. To the best of my knowledge, this paper is the first, other than the present work, to include more sophisticated syntactic features derived from full parses as part of its contextual information. The main point of interest for the authors is the notion of the importance of correct PP attachment. This relates to the issue of the adjunct/argument distinction: since arguments, fulfilling a complementation role, may depend closely on the lexical items to which they are attached, it is crucial that appropriate attachment is identified to ensure these cases are dealt with correctly. Given the stance taken on the importance of syntactic analysis, the results achieved by this method are of particular interest. To assess the real contribution of syntactic information, the paper attempts the task with and without the inclusion of PP attachment features (in the 'without' dataset, sequential information is relied on instead). A memory-based learning framework is used, achieving overall accuracy of up to 71% on L1 data; no experiments on L2 data are reported. Crucially, the inclusion of syntactic features is found to bring a positive improvement to the results, which is in agreement with what is expected to be found in the present research.

Related to this research is work by Eeg-Olofsson and Knutsson (2003), which is also concerned with preposition errors, albeit in L2 Swedish. Their approach differs from most of the other work presented here in that it relies on manually crafted error detection rules rather than the acquisition of correct models from L1 text. These rules analyse the morpho-syntactic level of the relevant phrases. Some of the problems discussed by the authors are also found in the present work, as detailed in Chapter 6: most notably, they observe that the presence of other errors in the data makes the application of the system's specialised knowledge more difficult. Evaluation is minimal; a recall score of around 25% is reported.

In this thesis, both preposition and determiner errors are approached using analogous methods, differing only in the choice of features for each component. It is not the only work to address more than one error type in this way; other recent work also discusses these two error-prone POS together. Lee and Seneff (2006), for example, are concerned with preposition and determiner errors, as well as others involving verb and noun forms and auxiliaries, in the context of an interactive dialogue system aimed at L2 English speakers. The central idea is that various insertion points for the relevant

POS are hypothesised within the input sentence, turning it into a lattice of alternatives which are then scored by a language model. Training and testing are carried out on transcriptions of call data from the Mercury flight domain, which, as well as being a limited domain, also differs from much of the other work discussed in being spoken rather than written in origin. Determiner insertion achieves precision of 86% and recall of 73%, while preposition insertion achieves 83% and 70% respectively. Given the differing premise of this work, which hypothesises and assesses insertion points rather than evaluating elements already present, it cannot be directly compared to DAPPER's results. However, it is noted that the authors find that the introduction of parsing improves results, in agreement with the conclusions reached by Lee and Knutsson (2008).

The work reported in Gamon et al. (2008) is another example of a model focusing on both prepositions and determiners. While the underlying principle of training an L1 model to detect erroneous use in L2 writing is the same as many others described in this section, the approach chosen is different. The authors train a decision tree on text from the Encarta encyclopedia and the Reuters news corpus; the feature set consists of several basic local features including lexical items and POS tags (see Section 3.1.4.1 for a more detailed comparison). Two classifiers are trained for each POS, one to determine whether there is a need for the POS in question, and the other to choose the most appropriate lexical item for that context, meaning the system can deal with errors of omission and redundancy as well as incorrect choices. The suggestions output by the classifier are then scored by a language model component, trained on the English Gigaword corpus (Graff 2003): if the classifier choice receives a significantly higher score than the student's choice, the classifier choice is given as a possible correction. Accuracy on L1 determiners is 86% and just under 65% on preposition data; these figures are 59% for determiners and 56% for prepositions on the L2 data task. In Sections 6.3 and 6.4, these results will be discussed in greater detail.

Finally, work on English spoken by Japanese learners, carried out by Izumi and colleagues (see e.g. Izumi et al. 2004), is described. This work differs from much of the research in this section in that the data used consists of transcriptions of spoken English rather than written material. The data is tagged for discourse features as well as errors; obviously error types peculiar to written language, such as spelling mistakes, will not be present. One of the aims of the work is to automatically detect errors of all three types – omission, insertion, erroneous word. The information used is similar to that found in much other work. For omission errors, this consists of

surrounding words, POS tag, lemmas of these words, combination features of these, and the first and last letters of the word immediately following the putative omission point. Missing and replacement errors, in addition, also include this information for the target word; the first and last letters refer to the target word itself. Compared to the present work, there is no syntactic dimension; semantics is captured only indirectly through the presence of surrounding lexical items. Similarly to Dapper, on the other hand, a maximum entropy classifier is used. The approach is tested on 13 error types covering a variety of POS, including prepositions and determiners. The classifier is trained on both error-tagged data and correct L1 data, to recognise both correct and incorrect instances. Figures for each error type are not reported, apart from determiner errors. They are 68% precision and 29% recall for determiner omission; 68% precision and 16% recall for determiner insertion/replacement; 70% precision and 20% recall for other omission; 22% precision and 6% recall for other insertion/replacement. Many of these figures are rather low, especially as regards replacement errors for other POS; as it is not known how many of these refer to prepositions, no firm comparisons can be drawn with the results presented in this thesis.

2.3.3 Models of preposition syntax and semantics

The work addressed in this section involves the collection of lexical, syntactic, and/or semantic information about prepositions' contexts. Using this information as the starting point for a database of preposition usage, to be used as a resource in other applications, can be easily envisaged, and indeed, as noted in Chapter 1, is one of the longer-term goals of this research. One such project has already been undertaken, namely PrepNet (Saint-Dizier 2005), whose aim is to create a resource cataloguing prepositions' syntactic and semantic behaviours and their usage.

Despite the availability of such a resource, it was decided to create a new model to ensure the information available is better attuned to the particular needs of the task; for example, at this stage it is not necessary to focus on the fine-grained semantic distinctions offered for polysemous prepositions, but only on more abstract contextual patterns. It would be interesting, however, to fully develop a similar resource to compare findings obtained by the different methods.

A somewhat different, but related undertaking is that of *A valency dictionary of English* (Herbst et al. 2004), which lists complementation patterns for 511 verbs, 274 nouns, and 544 adjectives on the basis of data from the COBUILD Bank of English; this includes PP complementation. The work is aimed at advanced learners of English,

among others, and is in some ways comparable to the present research in its reliance of corpus data to extract patterns rather than defining them *a priori*. Although this is a comprehensive and thorough resource, its usefulness is in part limited by the fact that it does not appear to exist in digital form; it would be interesting to compare the patterns and regularities in PP complementation found by the authors with those extracted by the system presented here, which has greater coverage of lexical items.

2.4 Conclusion

This chapter has provided an overview of the types of analyses carried out on L2 data and the problems and benefits of using this resource, as well as presenting several approaches to error correction in language, with a particular focus on prepositions and determiners. Throughout the discussion, the emergence of some shared concerns has been observed; these will be summarised in the rest of this section.

2.4.1 Data sources

There is little overlap in the choice of both L1 and L2 data used in testing and training for these tasks, which may limit the ease of replicating results. For L1 data, resources like the WSJ and Reuters are often chosen; however, this may not be the best option as they offer limited domain coverage and could lead to the acquisition of skewed models. As discussed in greater detail in Section 3.2.1, it was decided to avoid this risk by training on the British National Corpus instead, which includes several different kinds of texts, and may therefore lend itself better to generalisations. However, this does not guarantee similarity to L2 data.

As observed in the course of this chapter, there is also little overlap in the choice of L2 material used for testing, partly due to intellectual property restrictions on some of the data. Although arguably it is beneficial to develop applications on several different text collections, to avoid the danger of overfitting all models to a particular corpus, having a shared resource would allow more direct comparison of results.

2.4.2 Feature selection

Appropriate feature selection is an issue central to the work on this topic, with most papers identifying a broadly similar set of relevant characteristics. The main differences, as shown in more detail in Chapter 3, lie in the way these features are included in the model. The feature set developed shares core similarities with several others,

but is enhanced by the presence of syntactic and semantic analysis, which may prove a more successful way of capturing relevant characteristics of the context.

2.4.3 Evaluation

As regards results, it must first of all be observed that error detection appears to be a difficult task for all. Despite variations in methods, none of the work discussed reports a high sucess rate, and even those presenting a better than average precision score do so at the great expense of recall. It is therefore important to continue research on this task, to establish what particular elements of each method may be most helpful and understand how performance can be further improved.

Furthermore, a common concern regards the difficulty of satisfactory evaluation. Several factors may mask the ability of the systems developed. These include the presence of other errors which cannot be treated by a given system, and the fact that several different options may be correct in a particular context, not all of which may be identified by the system. These issues arise in the present work as well, and some ways to address them are proposed in Chapter 6.

In the remainder of this thesis, the points raised in this chapter will be addressed and discussed in greater detail. This begins with a description of the principles underlying the development of Dapper, in the next chapter.

Chapter 3

The L1 model: motivation and methodology

The success of DAPPER, the model presented here, rests on the claim that preposition and determiner usage, although challenging to describe, is not entirely idiosyncratic or unpredictable. Patterns of occurrence can be extracted from reliable L1 sources and used to train a machine learning classifier, so that it can assign a label to novel instances with confidence. A flowchart of the approach is shown in Figure 3.1. In this chapter, the procedure developed to train a classifier to recognise correct preposition and determiner usage is explained. It begins by discussing the features selected as relevant for each task and the motivation for these choices, before describing the various components of the system.

3.1 Describing the context

As stated in Chapter 1, the work proposed here is based on the assumption that preposition and determiner choice is governed by the interaction of a number of syntactic and semantic features. This is not a novel assumption: much of the work discussed in Chapter 2 also relies on contextual features. Grammars of English, too, although they do not explicitly use terms such as 'features', describe usage rules on the basis of the elements of a sentence surrounding the lexical items under discussion.

However, it is not always easy to determine what these features should be. Indeed, when asked for guidance, native speakers often find it difficult to come up with rules which can be explained in a few words. This section gives a detailed description of the features chosen as relevant for the task and motivates these choices, as well as comparing this feature set to others found in the literature. The actual procedure of

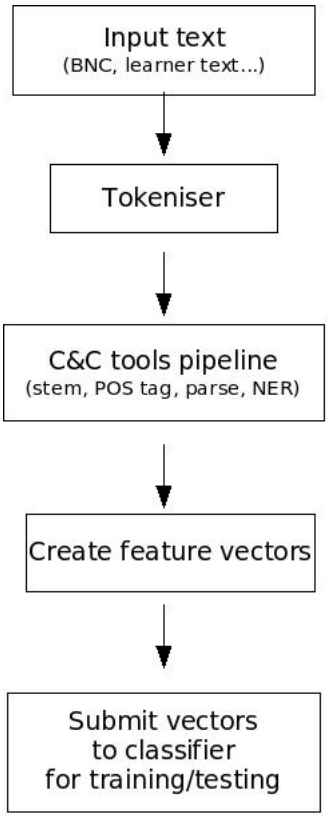

Figure 3.1: Schematic overview of the workflow described in this thesis

feature extraction will be explained in Section 3.2, while a detailed investigation of the contribution made by each feature is presented in Chapter 4.

3.1.1 The grammars' approach

The task of selecting a feature set begins by examining whether there are any grammars of the English language which contain rules and suggestions that can be usefully adapted as a feature set, or at least be used as the starting point for one. To this end, the treatment of prepositions and determiners in several resources aimed at both L1 and L2 speakers of English is considered.

3.1.1.1 Prepositions

Traditionally, the treatment of prepositions in grammar books is based on references to cognitive frameworks such as spatial, temporal, and so on, together with schematic illustrations, where the spatial sense is most often referred to as the 'prime' field from which all other metaphorical uses derive (see for example Tyler and Evans (2003), which is a study couched within a cognitive framework rather than a grammar

proper). Greenbaum and Quirk (1990), for example, differentiate among the various dimensions (one-, two-, three-dimensional) prepositions can refer to, and between direction vs. static position, as well as discussing several metaphorical uses. While the criteria discussed are certainly valid, there is a lengthy description which would be hard to distil into a set of features suitable for a machine learning task. Arguably, however, some of the concepts mentioned – staticity vs. movement, dimensionality – could be in some way captured by a feature set if semantic categories were included. For example, features could be introduced to distinguish between 'verbs of motion' and 'verbs of state'. The non-spatial uses yield little information that is specific enough to be included in a possible feature set, apart perhaps from the observations noted in their discussion of the temporal category, regarding usage rules for dates and times.

Huddleston and Pullum (2005) discuss general syntactic properties of prepositions, and contrast them to other POS. They then mention 'uses' of prepositions other than spatial or temporal, which they refer to as 'grammaticised' uses: cases where prepositions mark grammatical functions. Their examples include phrases such as NPs modified by ***of*** (*the death* ***of*** *the president*) and verbs, nouns, and adjectives that subcategorise for a particular preposition: *transfer money* ***to*** *him*, *request* ***for*** *assistance*, *keen* ***on*** *the idea*. While this text contains much useful discussion on when and how PPs are used, and the different forms they can take, there is virtually no information about how to infer preposition usage in any given context.

It seems, then, that grammars aimed primarily at native speakers do not consider it relevant to include guidelines on how to use specific prepositions in context. Texts aimed at non-native learners of the language, who often have to overcome the obstacle of significant divergences with their L1, may on the other hand be expected to be more explicit on the matter. Parrott (2000), written for teachers of L2 English, lists and discusses the various functions that prepositions can take on – spatial, temporal, logical, 'dependent' (i.e. grammatical), idiomatic – but, once again, more detailed guidelines are not offered.

The best sources of usage information are most likely to be guides to the English language designed for learners, such as Wood (1967), Sinclair (1991a), or Lindstromberg (1995). Here we find the kind of detailed explanations that are closest to what we could term features, descriptions of each preposition with examples of the types of words that can occur with them, both as objects and heads. For instance, the entry for ***from*** in Sinclair (1991a) states, among its twenty definitions:

> You use **from** when you say [...] where someone or something started off; where someone works; when giving the distance between two places; when you are talking about the beginning of a period of time or the first of a range of things; if something is made of a particular substance; when mentioning the cause of or reason for something; to indicate that something is being prevented or forbidden.

Entries in Wood (1967) are also typically rather lengthy. The one for **with**, for example, includes sixteen "chief uses" and seven "special uses", such as:

> [...] to express the notion of being accompanied by, or in the presence of; the idea of association or reciprocity between people or things; association or identification on the level of ideas or beliefs; to denote an instrument; to denote a cause or reason; to express attendant circumstances or state of affairs.

While these definitions and guidelines are very thorough and informative, they cannot be used directly to design a feature set. Rather, the main elements they indicate as being significant have to be extracted and ways of representing them through the NLP tools available have to be found.

Lindstromberg (1995) differs slightly from the other two texts in that, as discussed in the previous chapter, the author believes that there is a great degree of systematicity in preposition usage which learners of the language can acquire. His aim would suggest that the discussion in his volume could play a key role in the design of a feature set for the present task. In fact, we find a presentation of the topic based around a cognitive framework, together with some semantic rules which group lexical items into categories such as 'battle locations', 'verbs of looking', 'activity or event', and so on, noting for each category what preposition is most likely to be used. Such semantic information could be thought of as feature-like, although arguably it is very fine-grained and not as easily generalisable as might be desired.

3.1.1.2 Determiners

Unlike prepositions, determiners seem to have received more detailed treatment regarding their usage in the traditional grammar literature. Hill (1966) offers a good overview of several decades' worth of work on the topic, most of which in some form or other picks out as relevant factors familiarity, previous mention in discourse, countability, and whether a noun is proper or common – all of which come with at least

a few exceptions. The author's conclusion is that there is not "a great unanimity of opinion" on the matter, and "[previous explanations] do not add up to a clear explanation of the syntax and use of *the* in English" (Hill 1966:217). Greenbaum and Quirk (1990) devote almost twenty pages to the topic, which would be arduous to distil into a set of feature extraction rules. Uniqueness is mentioned as a key element in determiner choice, be it uniqueness in the local discourse or through shared knowledge of the world. It can also be derived from logical and grammatical factors such as modification by superlative adjectives and ordinals. Countability is also, of course, an important consideration, as well as proper names, which however present several exceptions. Furthermore, special usage rules are noted with regard to locatives, means of transport and of communication, time expressions, meals, and illnesses – a level of detail which may perhaps be captured with reference to semantic classes. There is little reliance here on syntactic features, the emphasis seemingly being on discourse and semantic aspects; a challenge for the present research could be to establish whether it is possible to identify usage patterns for determiners without having to rely on a semantic- and discourse-heavy analysis.

Strikingly, Huddleston and Pullum (2005) cover the topic of determiner use in less than four pages. Uniqueness of reference is the only criterion mentioned in deciding what type of determiner to use in a given NP. As for the learner-oriented texts, Parrott (2000:45) states that determiner choice is based on "a complex interaction of factors including meaning, shared knowledge, context and whether the noun is singular, plural, or uncountable" – indeed, the complex interaction the present work is attempting to model. General guidelines are then offered, mentioning number, countability, the presence of other determiners, and uniqueness of reference. There is also an attempt to give more practical, usage-based advice, such as using indefinites with 'naming things' (*this is a book*) and occupations (*I am a teacher*), or definites with relative clauses and some PPs. There are also further notes on particular items of a more specific nature, such as expressions of time and quantity, forms of entertainment and travel, institutions, and, again, meals. While some of the points raised by the author are valid, and sufficiently general to be included in a contextual model (such as PP or relative clause modification), in other respects his overview is too ad-hoc to be ported entirely into such an approach. However, this work is notable for its attempt to give practical examples of phrases showing how to use determiners.

Finally, Berry (1993) takes this approach one step further by devoting an entire short guide to determiner usage, in the same series as the one on prepositions (Sinclair 1991a). The topic is dealt with in a very detailed manner. The guide attempts to

cover most of the likely cases of determiner context, in terms of both syntactic and semantic features. As regards the former, we find, once again, mention of count and mass and reference, uniqueness, and shared knowledge; 'nouns with qualification' (PP, relative clauses, appositions); superlatives and other adjectives, among others. As for the semantic aspect, lexical items are grouped according to quite specific categories: 'media and communications', 'means of transport', 'musical instruments', 'illnesses', 'names of festivals', to name just a few. Again, while it may not be feasible to include all this information in a contextual model, this approach is further support for the notion that there is a need for a semantic dimension beyond the actual lexical item of the head noun in the model, such as named entity information, or perhaps WordNet classes (Fellbaum 1998). It is interesting to note that the two Collins guides to prepositions and determiners are corpus-based, having been developed with reference to a 25 million word corpus of English (the *Birmingham Collection of English Text*). In Chapter 5, a detailed analysis is given of the role played by individual features, and some generalisations on their co-occurrence preferences with particular items are formulated. These results could be compared to the advice included in the guides, to establish whether the use of different methods and sources has led to similar conclusions.

In theory, the texts discussed in this section, especially those on determiners, are sufficiently rich and detailed in information for a feature set to be constructed according to their advice, if one had lexical resources that were equally detailed and corresponded well to the semantic categories identified by the books. However, as well as being an extremely time-consuming approach, there is no guarantee that these authors have been comprehensive in their treatment of each possible case. It is preferred to automate the process of feature extraction to a degree, to ensure the feature vectors are free of any human bias (aside of course from that intrinsic to the choice of feature categories). By including contextual information at different levels, not just semantic type, the derived patterns will be more likely to be applicable to unseen instance. Of course, one can then return to the data extracted and *a posteriori* attempt a linguistic description of preposition usage in a similar spirit to the ones cited here.

In choosing these features, it is noted that several are specific to one or the other POS; however, there are some generally applicable principles that were followed in formulating a list of what to focus on. An important source, more so than the grammars, is the analysis of errors found in L2 writing. When looking at an error,

the identification of the elements of the sentence which were making that particular preposition or determiner choice sound awkward was attempted: these elements might then be identified as potentially relevant. For example, in *the training programme will start* ***at*** *August*, one could say that a wrong preposition is being used because months occur as the object of *in*, not *at*. Therefore, proper names of dates would be included in the preposition feature set. Similarly, in *this is caused by another contract with* ***the*** *another company*, the determiner error involving ***the*** is given by the presence of the quantifier *another*, so this element is added to the determiner feature set. Although this is not a guarantee of comprehensiveness, it is a good starting point on which to build.

Another important characteristic is that the features be easily extractable from texts, and therefore represented in some way in the output of standard NLP tools. This has been adhered to, as will be seen in Section 3.2. Finally, native speaker intuition was also relied on in some measure. It is possible that it has been erred on the side of caution, and more features than are helpful for the classifier have been included: this will be investigated in Chapter 4. In the sections that follow, the features for each POS will be described in detail. Tables 3.1 and 3.2 in Section 3.2 present this information in a summarised form.

3.1.2 Prepositions

3.1.2.1 Syntactic features

Among the syntactic features, one of the most important is the **lexical item the preposition is modifying, and its POS**. This can be one of noun (*the man* ***by*** *the door*), verb (*drive* ***to*** *London*), pronoun (*a photo of you* ***in*** *a hat*), adjective (*guilty* ***of*** *treason*), or adverb (*away* ***from*** *home*). For adjectives and adverbs, there is also a feature for the grade of the adjective or adverb – base, comparative, or superlative – as different forms of these items can show stronger collocational patterns with particular prepositions, for example superlative adjectives appear very often with ***in*** or ***of*** (*highest* ***in*** *salt*, *easiest* ***of*** *the assignments*). Knowing the POS of the word modified by the preposition is important as many prepositions tend to occur more or less often with different POS. For example, ***of*** occurs much more frequently with nouns and adjectives than with verbs, while ***from*** displays the opposite behaviour, appearing more often with verbs.

But what is the importance of the individual lexical item features? Are they really necessary? Not only does their presence cause the feature space to increase vastly,

but it would also appear to undermine the claim that preposition usage follows generalised patterns rather than being idiosyncratic and word-specific. The extent of the contribution made by these lexical items will be quantified in full in Chapter 4, where the effect of various feature combinations on accuracy will be analysed. However, for the time being it can be assumed that both positions are possible. Undeniably, some high-frequency lexical items have such strong collocational patterns with particular prepositions that their presence alone is likely to bias the classifier towards the correct choice. Examples of these might include words such as *give*, *guilty*, *discovery* – items whose semantic structure requires a preposition to be semantically complete. On the other hand, the aim DAPPER sets out to fulfil is to deal correctly with any context it might encounter, not just high-frequency terms. To this end, then, the extraction of more general patterns is essential, and it is to these non-lexical-item-dependent features that the system can turn when the lexical items present do not provide useful clues as to what the intended preposition might be. Furthermore, it is advisable not to rely too much on single lexical item-preposition combinations. While strong collocations do exist, they are not always exclusive: for example, you can *give something* ***for*** *an occasion*, as well as *giving it* ***to*** *someone*. In these cases, the broader context should assist in directing the classifier towards an appropriate correctness judgement, for example noting that the verb may be already modified by another PP.

The **preposition's object** (e.g. ***with*** *the spoon*, ***for*** *his coming*) is treated similarly, except that fewer possible POS are considered, namely only nouns and verbs[1]. Here, too, the issue of the contribution made by the lexical items is open to debate. It would appear that, in object position, the lexical item's role should be less important because there are fewer collocational restrictions; language is an open and creative system, and almost anything can be a prepositional object (*eat the pizza* ***with*** *the fork, prosciutto, the peppers, your hands, your mouth, the dead fly...*). It would then be of more use to know what patterns can occur rather than what particular lexical items. However, there are also certain classes of words where the specific items are central to preposition choice. For example, the semantic field of dates: we say ***in*** *June*, ***on*** *Monday*, ***at*** *2 o'clock*, despite the fact that these are all expressions referring to a point in time. In cases like these, then, having the extra information about the actual item present rather than just noting that it is a named entity denoting a time or date will be of crucial importance.

[1]Although not all grammars agree that verbs can in fact be objects of prepositions; indeed, arguably such *-ing* forms behave like nouns syntactically and should be considered deverbal nouns. However, they are tagged as verbs by the tagger, and so are referred to as verbs here for convenience.

Also included is a feature referring to cases where a preposition modifies **a verb or noun modified by more than one PP**, such as *travel **from** Naples **to** Rome* or *the book **of** poems **from** John*. This information may be useful to record because some prepositions, such as ones denoting spatial relations, are more likely than others to occur in these contexts. Similarly, there is a feature noting whether the modified verb appears with a direct object and is therefore transitive; again, some prepositions may be more likely than others to modify transitive verbs.

A further syntactic dimension of the context is captured by a feature which records the **grammatical relation** (GR) the preposition is in with its object and with the word it modifies, for example non-clausal and clausal modifier, indirect object, direct object. This is motivated by the observation that not all prepositions occur in all GRs, so that the presence of some of the less widespread ones might constitute a distinctive feature in classification. For example, ***for*** and ***to*** seem to be the only two prepositions to occur in a *ccomp* (saturated clausal complement) relation with their heads (*impossible **for***, *waiting **for***).

Finally, a simpler feature is also taken into account, namely the **POS of words in a three-word window** either side of the preposition, as a way of capturing some low-level local contextual information. This information does not require any syntactic processing, but only POS tagging; it will be informative to examine whether more basic processing is sufficient to achieve satisfactory results, or if the additional levels of processing such as parsing bring a significant improvement. The belief is that simple sequential information alone – that is, not derived from deeper analysis of the sentence – may not be sufficient to acquire occurrence patterns for prepositions, given the potential for long-distance dependencies. It can be argued that in a phrase such as *the girl in the photo **with** the long red hair* it is more informative to know that the preposition ***with***, for example, modifies the noun *girl* and has the complement *hair* rather than the fact that it is followed by a determiner and an adjective and preceded by a determiner and a noun with which it has no syntactic relation. However, if one looks more closely at the kind of information offered by the tags of even a relatively small tagset such as that of the Penn Treebank, it can be seen that other kinds of potentially useful information can be derived from it. For example, we can learn if the nouns in the immediate vicinity of the preposition are singular or plural (through their tags) and mass or count (through the presence or absence of determiners), thus adding another dimension of syntactic-semantic content. Similarly, the various verb tags carry information about tense and form (for example gerund, past participle), which can prove useful in distinguishing between prepositions.

Furthermore, in this way it is possible to capture other information about possible modifications and complements which are less frequent and therefore not included in the list of features regarding syntactic dependencies, but might still occasionally surface and prove informative. This includes other prepositions and numerals. As regards the prepositions, there are cases where a preposition has as its complement another prepositional phrase (e.g. *out of the box, next to the desk*). Having this information captured by the POS context may prove useful, since not every preposition is found in this type of relation – ***of*** and ***to***, for example, seem to occur in them more frequently than others. The Penn Treebank tagset includes the tag 'CD' for cardinal numbers; although the named entity recogniser (NER) features described in the next section already account for several occurrences of numbers – dates, times, currencies – noting the presence of a numeral may be useful in various ways. First of all it, too, gives some information about the type of noun, as a numeral followed by a noun obviously points to a count noun. Also, there are certain prepositions which are particularly frequently found in constructions involving numerals and certain types of nouns which can be identified by their WordNet categories – ***in*** *10 minutes,* ***for*** *2 days*, and so on – so adding this numeral feature could help identify the presence of such phrases.

3.1.2.2 Semantic features

In the previous section, it was noted that both syntactic context and individual lexical items play a role in determining preposition occurrence. Do semantic characteristics have anything to contribute? The reasoning behind this assumption would consider the benefit of generalising known patterns, whether lexicalised or more abstract, to unseen instances. For example, if *draw* ***with*** *a pencil* and *sketch* ***with*** *a pencil* are known to be correct phrases, when a verb semantically related to the previous two is encountered, such as *scribble*, it can be assumed that it will take the same preposition.

There are of course several different ways these semantic relations could be obtained. A simple one would be to create clusters of words from a large corpus, based on similarities in their co-occurrence patterns. Alternatively, an already existing resource could be used, such as **WordNet**, where lexical items are arranged and presented on the basis of the various relations linking them to each other. As it is believed that a rather broad level of semantic characterisation (e.g. 'verbs of motion', 'nouns denoting animals', and so on) should suffice, the latter resource was chosen, as will be described in more detail in Section 3.2.

The other major semantic component regards the presence of **named entities** (NEs), as both heads and objects of PPs. The six main MUC (Message Understanding Conferences) categories are used, namely Person, Location, Organisation, Time, Date, and Money. There are several reasons for including this feature. For example, if a 'money' named entity is the object of a preposition, that preposition is more likely to be one such as ***for*** or ***to*** (*it sold* ***for 100 pounds***, *the price dropped* ***to 30 dollars***). Times and dates occur most often as the object of ***at***, ***in***, and ***on***, though the differences within these make the presence of the actual lexical items involved indispensable, as discussed above. As for heads of phrases, a striking example is the category of organisations, which display a strong preference for prepositions such as ***of*** and ***for***, as in ***Association for*** *the Protection of Animals*, ***Institute of*** *Advanced Studies*, and so on.

Finally, a further feature takes into account other lexical properties of a large number of verbs, namely their **subcategorisation frames**. This gives information on whether they are transitive or intransitive, can take any PP complement, or only ones headed by specific prepositions such as *apply* (frame: np_ppto[2]) or *suffer* (frame: ppfrom[3]). The motivation for the inclusion of this feature is that if certain verbs are recorded as occuring only or prevalently with a particular preposition, the presence of this feature should provide a strong clue as to what the correct preposition should be. It is in fact debatable whether these constructions are simply verbs modified by prepositional phrases, or phrasal verbs, that is, verbs where the preposition's contribution to the meaning is non-compositional. This is an important point because the tools used might treat these two groups differently, for example tagging prepositions in phrasal verbs as particles rather than prepositions, in which case those instances would not be recorded by DAPPER at all, since only elements tagged as prepositions are recovered by the system. However, should this information be present, it is likely to be useful, so the feature is included.

3.1.3 Determiners

The feature set for determiners is somewhat smaller than that for the prepositions, since here what is being described is only the relation between a determiner and its noun or NP rather than between two entire phrases. The label to be assigned is one of three possible determiner classes, the choice of which depends on the properties of the noun modified by the determiner. Therefore, most of the features describe

[2]Meaning the verb takes an NP object followed by a PP headed by ***to***.
[3]Meaning the verb is intransitive and is followed by a PP headed by ***from***.

characteristics of the head noun and the phrase it is found in. The question being asked is: given a NP with these properties, is a determiner required here, and if so, which one?

3.1.3.1 Semantic features

Naturally, the key element in the NP context is the head noun, that is, the noun in the NP which may or may not require the determiner. Several features relate to intrinsic properties of the noun, which is why semantic features are discussed first. In fact, the primacy of semantic over syntactic features here is further evidence of the difference between determiners and prepositions: for the latter, being functional elements, the syntactic dimension appeared more important.

The first feature is, of course, the **lexical item** itself; **its number**, equally important, is encoded in a separate feature through its POS tag. This allows the classifier to make generalisations about how often a particular lexical item occurs in the training data with each determiner class. The number is also essential as it restricts the field of available options. For example, while a singular count noun may in theory occur either with ***the***, ***a***, or no determiner, for a singular mass or plural noun the choices are limited to two cases only, ***the*** or ***null***. Nevertheless, it may still be quite challenging, with plural nouns, to distinguish between cases where the definite article is needed and those where none is required – cf. *boys like football* vs. ***the*** *boys like football*, which are both grammatical, and only differ in terms of whether one is making a general statement about boys or referring to a particular group of them. This depends on aspects of the discourse whose scope is broader than the individual NP or sentence; they will be discussed further in the next section.

The discussion above introduced another aspect of the noun, namely whether it is a **count or mass noun**; this distinction refers to its countability, i.e. whether the noun can be enumerated (cf. *a book, two books* vs. **a weather, two weathers*). Mass nouns do not as a rule occur in the plural, and cannot, as seen in the previous example, be preceded by the indefinite determiner ***a***. In reality, very few mass nouns are strictly mass nouns only and there are several idiomatic uses where they are found with an indefinite determiner, for example *I would like* ***a*** *water with ice and lemon*. However, these are not generally their primary use, and there are many other mass nouns which are never used in this way (e.g. *information*), so the inclusion of this feature is likely to play an important role.

Of course, it is possible that a particular noun occurring in testing has not been previously encountered in training, and there is no information about its countability

available to the classifier. For this reason, it is important that DAPPER rely not just on features peculiar to individual lexical items, but also on more generalised contextual elements, both semantic and syntactic ones.

As to the former, a key feature is whether the noun is a proper noun or **named entity** (following the categories listed in Section 3.1.2.2). This feature's importance derives from the fact that there are several important generalisations that can be made about determiner occurrence with proper nouns. Firstly, in English proper names of people do not, as a rule, require a determiner: one does not say **the Mary went to the store*. Occasionally one does find occurrences with an indefinite article, as in *she was looking for* ***a*** *John Smith, but she had the wrong number*, but this is rarer, idiomatic usage. Locations, too, tend to have no determiner in English, unlike, for example, several Romance and Slavic languages: *Naples*, *Italy* rather than *****the*** *Naples*, *****the*** *Italy*. This rule, however, has many exceptions in names such as ***the*** *United Kingdom*, ***the*** *United States*, ***the*** *Channel Islands*, and so on – in other words, those proper names where the head noun is also a common noun. These are the minority; their occurrence as lexical item features is expected to override any information contributed by the named entity feature.

Organisation proper names, on the other hand, display a lot more variation which may make their identification less informative for the classifier's purposes. This is found within the same type of organisation – for example ***the*** *BBC* but not *****the*** *ITV* – and even within the same entity itself – cf. ***the*** *University of Oxford* vs. *****the*** *Oxford University*. More straightforward are the usage rules for NPs denoting currencies. Usually these occur without a determiner (e.g. *This costs 300 pounds*, not **This costs* ***the*** *300 pounds*)[4], so the inclusion of a feature noting their presence should prove useful.

The final two named entity categories are time and date. Although there are occasional mistakes in the NER with regard as to what falls under each category, generally years, days of the week, and months are labelled as 'date', while terms such as *midday*, *afternoon*, *the Sixties* are classified as 'time'. It is important for the classifier to have access to this information because English, unlike several other languages (e.g. Italian), does not regularly use a determiner before date terms. So one says *in 1989*, not **in* ***the*** *1989*, or *Monday is the best day of the week*, not *****the*** *Monday is...*, and so on[5]. Conversely – and perhaps confusingly – some nouns

[4]But cf. *Here are* ***the*** *300 pounds I owe you*.

[5]However, there are cases where a determiner is used, such as ***the*** *Monday after next*, ***the*** *Christmas before last*; these represent more marked, less frequent usage.

denoting times, but not all, do require a determiner: *in* ***the*** *morning*, *in* ***the*** *afternoon*, but *at sunrise*, *at midday*. Therefore, while the NE tag identifying dates will offer a valid clue as to the need for a determiner, in these latter cases the actual lexical item together with other features, if necessary, will be required for the correct labeling of the instance.

A further semantic dimension consists of the identification of broad similarities among nouns by relying on **WordNet** data (see Section 3.2). The aim of this feature category is to complement the mass/count feature and add a further level of semantic distinction. However, as will be discussed in that section, there may be some drawbacks to the particular approach taken to make use of this data.

Finally, a further typical distinction made for nouns is that between abstract (e.g. *imagination*) and concrete (e.g. *stone*) nouns. This has some overlap with, but is not identical to, the mass/count distinction: many abstract nouns are mass nouns, such as *imagination, information, love*, but so are many concrete ones, such as *milk, bread, flour*. This unpredictability, together with the fact that there seems to be no direct correlation with determiner choice, means that this particular aspect of noun semantics is not included in the feature set. Arguably, the broad distinctions it would capture are already present in the WordNet classes.

3.1.3.2 Syntactic features

As already anticipated, one of the most important factors in determiner choice is the discourse dimension. Whether a noun requires a determiner, and which one, most often depends on its novelty within the discourse, i.e. if it is being introduced for the first time or not. Generally, unless there are other reasons to assume the intended audience is familiar with it or it is somehow unique (e.g. a public figure – ***the*** *Queen*, a well-known event – ***the*** *summer solstice*), a noun will occur with ***a*** the first time it is mentioned, and thereafter appear with ***the***. This implies that in developing models of determiner occurrence, one should look beyond the level of the individual sentence to the wider discourse unit of a paragraph, or indeed a whole text, taking account of phenomena such as reference and anaphora resolution – something which is not required for functional POS such as prepositions.

However, moving to a broader discourse unit may require a significant processing effort, both in terms of time and resources. As shown below, there are in fact a number of other aspects of the local context which are likely to play a role in determiner choice. Therefore, only syntactic and semantic features are relied on in training the classifier,

not just for reasons of economy, but also to investigate to what extent one can forego the discourse element in this task.

Several syntactic features can be extracted from the context at sentence level; for example, **adjectival modification**, especially with regard to the adjective's grade – base, comparative, or superlative. This can be an informative feature, especially in the latter case: it would be unusual to find an indefinite determiner with a noun modified by a superlative adjective (cf. ***the*** *biggest box* vs. *****a*** *biggest box*).

Whether the noun is modified by a **predeterminer** is also noted. This includes lexical items such as *such*, *all*, *both*, *half*, and so on. As the name suggests, these generally occur only before a determiner, in phrases such as *all* ***the*** *children* or *half* ***an*** *apple*. Therefore, their presence is an important indicator of the need for a determiner. However, it must be observed that phrases such as *both girls* or *all children* are also grammatically correct: in these cases, which the tagger would probably tag as adjectives, the classifier would have to draw on other contextual features to support its decision-making. Yet again, it is clear that there is no single 'silver bullet' feature governing these choices, but rather a subtle interaction of several different elements is at work.

In addition to adjectives and predeterminers, a noun may also be **modified by a numeral or a possessive**, both of which have an effect on the need for a determiner. With regard to numerals, there are two possibilities, ordinals (*first, second, third...*) and cardinals (*one, two, three...*). These are tagged differently by the POS tagger: the former are included under adjectives, the latter with the tag 'CD'. The fact that ordinals are grouped together with other adjectives is not a problem, because they will appear as a feature nonetheless, as 'adjective modifying the noun'. It is useful to record their presence in the NP because singular nouns modified by ordinals generally require a determiner: cf. for example ***the*** *first time I saw him*, not **first time I saw him*. Furthermore, this determiner is much more often ***the*** than ***a***. As for cardinals, they are especially useful in singular nouns, in blocking the occurrence of the indefinite article – cf. *he saw one person*, not **he saw* ***a*** *one person*.

Possessives refer to the possessive pronouns in phrases such as *his dog*, *their house*, and so on. In English, possessives occupy the same slot as determiners, making the two mutually exclusive. One does not say *****the*** *my house is on sale*, for instance. This can make their presence in the feature set extremely informative, as modification by a possessive automatically rules out the possibility of a determiner occurring with that noun[6].

[6]This property of English and other Germanic languages is in fact among those which cause the

Another element of the NP believed to contribute useful information is the **presence of PPs**. This is represented by four separate features, two for those cases where the head noun is the object of a preposition, and two for those where it is modified by a preposition. In both cases, one feature notes the presence of such a relation, and the other notes what the preposition involved is. Nouns in object position can occur equally plausibly both with and without a determiner, depending on the wider context: cf. *she danced on tables*, *she danced on* ***the*** *tables*; *he covered her with flowers*, *he covered her with* ***the*** *flowers*, so it may be that this feature will be found to be not very informative[7]. On the other hand, nouns modified by a preposition – especially ***of*** – are generally found to require the definite article, for example ***the*** *object of a preposition*. This may be due to the fact that the presence of such PPs restricts the field of possible referents for a NP, making it more likely to require a definite determiner.

Similarly, the presence of a **modifying relative clause** may also play a role, by occurring in NPs which show a preference for the definite determiner: cf. ***the*** *man whom we saw*, ***the*** *car that we like*. Analogously to the PP modification examples, these relative clauses usually have the function of singling out a particular person or thing, making it a unique referent. NPs such as *a girl that I know can fly* are of course grammatical, too, but appear to be less frequent. It is also hypothesised that a contributing factor to determiner occurrence is whether the noun or NP is part of an **'existential' or 'there' phrase**, such as *there is a lovely view from here*, *there is a girl here to see you*. Singular nouns in these phrases are more likely to occur with the indefinite article, as they are typically used to introduce a topic for the first time, though experimental data will be needed to clarify the validity of this claim.

Finally, as well as the features described above, which rely on some degree of syntactic and semantic analysis, also for the determiners, as for the prepositions, there is the some lower-level contextual information in the form of **POS tag sequences**. Again, a $\pm$ 3 word window around the determiner is considered. Where the head noun has a null determiner, the head noun is the centre of the window. Although most of this information will have been captured by the other features, it is nevertheless useful, as discussed in Section 3.1.2.1, to refer to the immediate sequential context as well, for those rarer elements of the context that might otherwise go undetected.

---

most problems for learners whose L1 is a Romance language, as in that language group, determiners and possessives do, and indeed must, co-occur: for example, in Italian one says *la mia casa* ('my house') rather than just **mia casa*.

[7] Though there are also distributional differences in PPs such as *by bus* – *on* ***the*** *bus*, where knowledge of the prepositional information might prove determining (cf. also Nagata et al. (2006a)).

3.1.4 Feature set comparisons

As noted in Chapter 2, there are other ongoing research efforts on this task. In this section, the choice of features used in this thesis will be compared with others in the literature to establish similarities and divergences.

3.1.4.1 Prepositions

In their work, Chodorow et al. (Chodorow et al. 2007; Tetreault and Chodorow 2008a,b) use 25 contextual features for their preposition model, mostly of a local and syntactic nature. They do not carry out a full parse of the sentence, but only stemming, POS tagging, and NP and VP chunking. As they rely on linear sequences of chunks rather than parsing, most of the features are of the type 'preceding noun', 'lemma of following verb', the POS tags of these items, and so on. Lexical item bi- and tri-grams are included, too. 'Combination features' are also introduced, where some of this information is merged into a single feature. Some data extracted from the Google n-gram corpus[8], referring to the most frequently seen sequences of nouns and verbs with the target preposition, is added as well. These two additions bring the total of contextual predicates to 41. Despite the lack of parsing or any other deeper syntactic analysis, it is likely that the overlap between the type of information captured by these features and the POS/lemmas involved with the preposition will be high, especially where the object is concerned. However, it is also possible that some instances of more complex PP attachment will go undetected.

Gamon et al. (2008) also rely on lower-level processing only, namely POS tagging and n-gram sequences. They consider a six-token window either side of the target position (a potential site for preposition occurrence), determined on the basis of POS tag sequences, and within this window they take into account the POS and lexical items present. Here, too, then, the emphasis is on linear sequences of words and POS tags, presumably on the assumption that a PP will not be found very far from the other phrases it is linked to. Certainly, while this approach requires less processing of the text, it might not allow for the extraction of more abstract patterns.

The approach of Lee and Knutsson (2008), as mentioned in the discussion of their work in Chapter 2, is most closely comparable to the one of this thesis in its intention of assessing the value of including PP attachment as derived from parsing. They create two feature sets for this purpose, one with parsing information, the other

[8]The Web 1T corpus, also known as the Google n-gram Corpus (Brants and Franz 2006), is a collection of 1 trillion words collected from the web, arranged in n-grams from 1 to 5.

without. In both sets, the features include the head of the PP phrase (derived from proximity or from parser output) and the head of the object of the preposition.

Overall, the main differences between the feature set used for DAPPER and those discussed above consist in the fact that the other models do not incorporate any semantic information, nor, with the partial exclusion of Lee and Knutsson (2008), any deeper syntactic information such as the relations entered into by the various elements of the sentence. While all approaches are focused on the nouns, verbs, and other lexical items in a preposition's context, here this information is extracted on the basis of syntactic relations, while others choose to rely on linear ordering and chunking only. A comparison of the results obtained by DAPPER and these models will therefore provide useful information with regard to whether a more in-depth approach to feature extraction is beneficial to this task, although these conclusions will be limited by the use of different data sources.

3.1.4.2 Determiners

As seen in Chapter 2, the body of work on determiner use is somewhat larger than that on prepositions; it is also more uniform, in the sense that there seems to be a high degree of agreement as to what constitute relevant and informative features for this task. In the literature, Knight and Chander (1994) is one of the very few papers to debate the merits of relying on hand-crafted rules found in grammar books for indications as to what to include in a feature set. Their conclusion is similar to the stance taken here, that is, that such rules and guidelines are often difficult to put into practice, riddled with exceptions, and probably not exhaustive, making a data-driven, automated approach preferable. The scope of their analysis is limited to the NP, within which they extract as features the following: head noun, presence of premodifiers (*next*, *same*) and superlatives, number, countability, and information on the lexical items surrounding the noun, including their POS. There is no attempt to integrate semantic, discourse, or deeper syntactic analysis here.

Minnen et al. (2000), on the other hand, add some more sophisticated features to a set which also includes expected items such as the head noun[9], its POS, whether it is count or mass (as listed in the ALT-JE Japanese-to-English translation system), and the presence of other determiners. Their other features refer to the wider context the NP occurs in, such as the functional tag of the NP's head and of the category embedding it, as well as the category itself; these labels are taken from the Penn

[9]This is identified as the rightmost noun in the NP.

Treebank. Functional tags are syntactic categories such as 'adverbials', 'dative object', 'temporal phrases', while the category of the constituent embedding the NP refers to syntactic phrases such as PPs. Finally, there is also a semantic aspect to their feature set, as head nouns are assigned to a semantic class, also according to information found in ALT-JE. The choice to include semantic information, as well as the syntactic relations the NP is involved in, make this feature set very similar to the one of the present work.

Lee (2004b) is another example of an approach which incorporates both semantic and syntactic features. Most of the features are the same as those used by others: the head noun, its number, whether it is a common or proper noun, the category of the parent node (e.g. PP), whether other non-article determiners are present, and the lexical items before and after the head within the NP, together with their POS. Notably, this work does not consider countability as a possible feature. It does have, however, a feature based on WordNet hypernyms and a discourse feature noting whether the head noun has been mentioned before (within a five-sentence scope). Apart from the discourse element, here, too, there are strong similarities with the feature set of the present work.

Some more recent approaches, on the other hand, choose not to include semantics. Han et al. (2006) POS-tag and chunk their data and extract features from the local context of the NP: the words within it, two words preceding it, the word following it, and the POS tags of all the above. Additionally, the head noun's countability is also a feature, although this is determined heuristically rather than by reference to any external sources. Finally, Gamon et al. (2008) use the same approach for determiners as they do for prepositions, namely extracting features based on lexical items and POS tags, while Yi et al. (2008) do away with corpus-based extraction and *a priori* feature selection altogether, and simply query the web to determine whether a given sequence is likely to be legitimate or not.

It would appear from this brief overview that there is broad agreement as to what constitute key elements in a determiner's context, and there is much overlap between the feature sets of the various studies discussed. The one used for DAPPER, too, is not very different in content, even if some of the information is captured in a different way. For example, while in the present work it is noted explicitly whether the head noun is modified by an adjective or a predeterminer (extracting this information from the parser output), others will gain this information by virtue of recording the presence of these lexical items and tags within the NP. As already noted in the discussion on the preposition features, the present approach may seem more processing-intensive

than that of others, despite the underlying similarities. This makes the analysis of differences and similarities of the results interesting, as it is important to establish whether such further processing is necessary for a satisfactory performance.

3.2 Methodology

Having established what features are to be extracted from the data, the discussion now turns to what resources to choose as data sources and what tools to use to create the instances that the classifier will be trained on.

3.2.1 Data

As mentioned in Chapter 1, in developing an NLP application ultimately aimed at learners of English some consideration must be given to the variety of English that will be taken as 'standard' or 'correct'. Here, it is assumed that the L2 texts used as a source are aiming towards correct British English, since they were written in the context of British-based exams; therefore, British English has been chosen as the reference variety for correct usage patterns. There are several other criteria to be considered, too. For robust generalisations to be made, a classifier needs a large number of training instances. Furthermore, to ensure that the application learns grammatically correct usage patterns, it is of course necessary for the data source to be reliable in this sense, that is, to contain texts which are believed to be grammatically correct. Therefore, the source must have been compiled according to strict editorial criteria, rather than just by collating data from the web, for example. A final requirement is that of diversity: if the patterns extracted are to be comprehensive and representative of the chosen variety of English, it is important that the data come from several different text types. This is because there may be certain PP constructions that only occur in one sub-genre, for example financial writing, and limiting the study to one such text type risks skewing the acquisition towards patterns typical only of one type of text.

The British National Corpus (BNC, Burnard 2000) fulfils all the requirements discussed above. Its size of 100 million words ensures that there is a sufficient quantity of data for each of the classes of interest. The quality of the texts is guaranteed by the careful planning that has gone into the creation of the corpus and the selection of texts to be included. The criteria followed also mean that there is a good representation of several different text types, such as newspaper and journal articles, academic and

fictional writing, letters, and essays, which should ensure that the findings of this research are not restricted to one particular sub-genre of the English language.

3.2.2 The classifier

The basic task it is proposed to solve is to model the probability of a particular class label being assigned to a given context:

$$P(\textit{preposition or determiner} \mid \textit{context})$$

A first approach to the prepositions task considered using a binary voted perceptron algorithm (Freund and Schapire 1999), where lexical items were not included in the feature set. Several binary classifiers, one for each target preposition, were trained. The only question these can answer is whether a given feature vector should or should not be assigned the label of a particular preposition. While this is not very useful for the ultimate goal of this research, it is a good first test for the validity of a machine learning-based approach, as it can show whether contextual representations in the form of feature vectors are an informative means of describing preposition occurrence and distinguishing between 'target preposition' and 'everything else'. Preliminary results on this type of task, on just five prepositions, yielded accuracies ranging from 69% to 84% (av. 77%; cf. De Felice and Pulman (2007)). These figures suggest that a machine learning-based approach to the contextual representation of prepositions is indeed viable and likely to give positive results, so the next step is to identify a classifier with the appropriate characteristics for the task.

The context of the target item can consist of several different features, including any number of varying lexical items. Furthermore, no assumptions can be made about the independence of the various features from each other; on the contrary, several are likely to be closely related. This is especially evident for the determiner set, where along with the presence of a named entity feature, for example, we might find the noun's tag as NNP, which encodes the fact it is a proper noun.

Maximum entropy algorithms (Ratnaparkhi 1998) are often used in NLP tasks because they meet all the requirements discussed above. No assumptions about feature independence are made; furthermore, a large number of heterogeneous information sources can be included as features, and if there are any interdependencies among them, these will be taken into account during training. This fact, together with the flexibility to include complex features, as language-based features often are, have made maximum entropy a very popular approach for a variety of NLP tasks (POS tagging,

parsing, document classification, to name but a few), and one that is well-suited to the task.

There are many freely available implementations of maximum entropy-based classifiers. In this work, a standard algorithm developed by James Curran has been used[10]. The classifier is accessed through a simple command line interface and causes minimal overheads in installation and usage.

3.2.2.1 Terminology

Throughout this thesis, several terms are referred to in relation to the use of the maximum entropy model. These will be defined more precisely here, following the approach used in Ratnaparkhi (1998).

So far, the term 'feature' has been informally used to describe those elements of the contexts of prepositions and determiners – the set of possible classes to be predicted – which have been hypothesised to be salient for the prediction of a particular class. These include observations such as 'preposition modifies a noun' or 'head noun is modified by an adjective'. In the maximum entropy literature, these observations are more properly defined contextual predicates. A ***contextual predicate*** captures the information present in a given context. For example, the observation that the preposition modifies a noun could be encoded by the contextual predicate *ModNoun*, which takes on a value of true.

While contextual predicates record the properties of the context of the target class, they do not carry any information about their usefulness in class prediction: this is encoded by features and the weights assigned to the latter. A ***feature*** is a binary function[11] which relates the co-occurrence of a class prediction and a contextual predicate. In the example below, a is an example of a contextual predicate and b is the class label to be predicted. In other words, the feature returns 1 only if the contextual predicate `ModNoun=gift` is found to occur with the preposition *for*.

$$f_i(a,b) = \begin{cases} 1 & \text{if ModNoun} = \textit{gift} \ \& \ \text{b} = \textit{for} \\ 0 & \text{otherwise} \end{cases}$$

The same contextual predicate may occur with several different classes: for example, many different prepositions can modify the same noun. It follows that not all contextual predicates, and by extension the features associated with them, will be

[10]James Curran's support in granting me access to this program is gratefully acknowledged.

[11]Although it is also possible to have non-binary features. For example, in maximum entropy parsing models features are integer-valued functions which count the number of times some pattern occurs in a parse. In principle features can be real-valued.

of equal importance in predicting a class. The extent to which features play a role in making the class prediction is represented by the ***weight*** assigned to a feature; weight assignment occurs in the training stage of a model.

The role of the weights in class prediction becomes clear if we consider the formula of the maximum entropy model:

$$p(a \mid b) = \frac{1}{Z(b)} \prod_{j=1}^{k} \alpha_j^{f_j(a,b)}$$

Here, $p(a \mid b)$ is the probability of the class label being a given the context b, k is the number of features, α_j is the weight assigned to each feature, and *Z(b)* is a normalising constant to ensure that a probability distribution is obtained. For each $f_j(a, b) = 1$, its weight α_j will be included in the product which gives the probability distribution for that class. This formalises the intuition that the presence in a feature vector of several features associated with a particular class increases the probability of that class being the correct one for the feature vector, as the product of the weights will be greater.

In the context of this thesis, we can say that:

– the elements noted informally in the previous section (noun is singular, preposition modifies transitive verb, and so on) are represented by *contextual predicates*, of the form (for example) `HeadNounSing`, `VerbModTrans`, `HeadNoun=x` (where x is a lexical item), and so on.

– the co-occurrence of such contextual predicates and a particular class is encoded by a feature, for example the presence of the contextual predicate `VerbModTrans` when the class label is *for*. This feature is assigned a weight.

– the number of weights to be learned, then, depends on the number of features, which in turn depends on the number of contextual predicates observed in the data for each class.

Below is a simplified example to illustrate these points. For this hypothetical set of training data consisting of the following two feature vectors, the relevant information is listed.

```
   for ModNoun NounMod:gift ObjNoun NounObj:sister
to ModVerb VerbMod:give ObjNoun NounObj:sister
```

- *contextual predicates*: ModNoun, NounMod:gift, ObjNoun, NounObj:sister, ModVerb, VerbMod:give

- *features*: (for, ModNoun), (for, NounMod:gift), (to, ModVerb), (to, VerbMod:give), and so on
- *weights*: 8 to be assigned; had the feature vectors been for the same class, only 6

In the remainder of this chapter, the procedure used to create the feature vectors is explained, and detailed examples of some of these feature vectors are presented.

3.2.3 Tools used

The ultimate output of the processing is to consist of feature vectors – class labels + contextual predicates – to be used in developing and training the classifier. The previous section discussed what types of contextual predicates are to be included for both prepositions and determiners. In this section, the tools used to create the feature vector instances will be briefly outlined.

3.2.3.1 The C&C tools

The first main step is to run the tokenised BNC text through the C&C tools pipeline (Curran et al. 2007). This toolset includes the CCG parser (Clark and Curran 2007), which can output RASP-style GRs (see e.g. Briscoe et al. 2006) as well as CCG structures, a POS tagger, a morphological analyser (morpha, Minnen et al. 2001), and a named entity recogniser (Curran and Clark 2003b). The various components of the pipeline allow the identification and extraction of all the contextual predicates of interest, from POS tags to syntactic dependencies and named entities. It is to be noted that the POS tagger uses the Penn Treebank tagset, which is smaller than others such as CLAWS[12]. However, it is believed that a smaller tagset will help in highlighting potential similarities in general patterns, which could be lost if a more highly differentiated tagset were used.

There are several robust parsers freely available to the NLP community, such as RASP (Briscoe et al. 2006), the Stanford parser (see e.g. Klein and Manning 2003), the Collins parser (Collins 1999), and the Charniak parser (Charniak 2000). The CCG parser, as Clark and Curran (2007) describe in more detail, is found to be consistently robust and efficient, as well as reliable on phenomena such as long-distance dependencies and coordination. This is especially relevant for the domain

[12]The Penn Treebank uses 45 against the over 160 tags in CLAWS 8.

of PP attachment, where dependencies can indeed be very distant from each other[13]; cf. *I saw the man with the red hat on his foot in the garden*, where the last PP is attached to a constituent (*saw*) which occurs much earlier in the sentence. If reliable usage patterns are to be derived, it is crucial that the parser be performing correct PP attachment.

The choice of using a parser rather than relying only on word sequences or chunking is motivated by the desire to ensure the correct elements are linked to each other, especially in the preposition component. As seen in the example in the previous paragraph, a PP may be sequentially at some distance from the item it is modifying; a parser can capture these relations in a way that simply noting a short n-gram sequence cannot. The same can apply to a preposition's object, if for example it is modified by several adjectives. In *we went to a grim, dreary, dark, and poky small town*, there are several words intervening between the preposition ***to*** and its object *town*, such that only a 7-gram might capture their relation, which risks going undetected. Therefore, it is believed that including full parsing is a valuable addition to the method presented.

3.2.3.2 Creating the feature vectors

Two separate Python scripts have been written to create the preposition and determiner feature vectors respectively. The basic structure is the same for both, and can be schematised as follows:

```
for each sentence in parsed text:
    if target POS found:
       create feature vector with target POS as class
       for each target contextual predicate:
            if contextual predicate present:
               update feature vector
```

The procedure is very simple, but lies at the heart of the successful implementation of DAPPER. In Section 3.1, the contextual predicates chosen and the motivation for their selection was discussed. Tables 3.1 and 3.2 summarise all the contextual predicates and briefly note through which tools the script recovers them. Where more than one possibility is given, these are intended to be mutually exclusive; x indicates any lexical item. A detailed example follows in the next section.

[13] F-scores for accuracy on PP attachment are around 85% for attachment to nouns and 71% to verbs.

Contextual predicate	Information captured	Source
HeadNoun:x	lexical item of head noun	constituent phrases
HeadNTag:NN/NNS/NNP/NNPS	number of head noun	POS tagger
NounType:count/mass/either	head noun countability	stemmer, external source
NEnoun:I-DAT/I-PER/I-LOC/I-TIM/I-ORG/I-MON	named entity and what kind	NER
WNet_Class_N:n	WordNet category	stemmer, external source
ModbyPrep	head noun modified by preposition	GRs
ModbyPrep:x	preposition modifying head noun	GRs
ObjofPrep	head noun object of preposition	GRs
ObjofPrep:x	preposition head noun is object of	GRs
ModbyAdj:x	head noun modified by adjective	POS tag, constituent phrases
AdjTag:JJ/JJR/JJS	grade of adjective	POS tagger
3left:TAG	POS tag of 3 words left of class label	POS tagger
2left:TAG	POS tag of 2 words left of class label	POS tagger
1left:TAG	POS tag of 1 word left of class label	POS tagger
3right:TAG	POS tag of 3 words right of class label	POS tagger
2right:TAG	POS tag of 2 words right of class label	POS tagger
1right:TAG	POS tag of 1 word right of class label	POS tagger
ModbyRelative	relative clause modification	GRs
ModbyQuantifier	presence of quantifier	POS tagger, GRs
ModbyPossessive	presence of possessive	POS tagger, GRs
ModbyCardinal	presence of cardinal	POS tagger, GRs
ExisThere	in 'existential there' phrase	constituent phrases

Table 3.1: Determiner contextual predicates

Contextual predicate	Information captured	Source
mod_noun	prep modifies noun	GRs, POS tagger
ModNoun:x	noun modified	GRs, stemmer
WN_Class_ModN:n	WordNet category of noun modified	stemmer, external source
NE_ModN:DAT/PER/LOC/TIM/ORG/MON	named entity and what kind	NER
ModN_morethanone	noun modified by more than one PP	GRs
mod_verb	prep modifies verb	GRs, POS tagger
ModVerb:x	verb modified	GRs, stemmer
WN_Class_ModV:n	WordNet category of verb modified	stemmer, external source
Subcat_frame:n	subcat frame of verb modified	stemmer, external source
ModV_morethanone	verb modified by more than one PP	GRs
ModV_trans	verb modified is transitive	GRs
mod_adj	prep modifies adjective	GRs, POS tagger
ModAdj:x	adjective modified	GRs, stemmer
ModAdj_type:JJ/JJR/JJS	grade of adjective	GRs, POS tagger
mod_adv	prep modifies adverb	GRs, POS tagger
ModAdv:x	adverb modified	GRs, stemmer
ModAdv_type:RB/RBR/RBS	grade of adverb	GRs, POS tagger
mod_pron	prep modifies pronoun	GRs, POS tagger
ModPron:TAG	tag of pronoun modified	GRs, POS tagger
ModGR_type:X	GR of prep & modified item	GRs
obj_noun	prep object is noun	GRs, POS tagger
ObjNoun:x	object noun	GRs, stemmer
WN_Class_ObjN:n	WordNet category of object noun	stemmer, external source
NE_ObjN:DAT/PER/LOC/TIM/ORG/MON	named entity and what kind	NER
obj_verb	prep object is verb	GRs, POS tagger
ObjVerb:x	object verb	GRs, stemmer
WN_Class_ObjV:n	WordNet category of object verb	stemmer, external source
obj_pron	prep object is pronoun	GRs, POS tagger
ObjGR_type:X	GR of prep & object	GRs
3left:TAG	POS tag of 3 words left of class label	POS tagger
2left:TAG	POS tag of 2 words left of class label	POS tagger
1left:TAG	POS tag of 1 word left of class label	POS tagger
3right:TAG	POS tag of 3 words right of class label	POS tagger
2right:TAG	POS tag of 2 words right of class label	POS tagger
1right:TAG	POS tag of 1 word right of class label	POS tagger

Table 3.2: Preposition contextual predicates

The tables mention some external sources of information which are not part of the C&C tools, used to include contextual predicates related to the WordNet classes, verb subcategorisation frames, and noun types (count or mass). The WordNet information comes from the 40 WordNet lexicographer files for nouns and verbs (25 and 15 classes respectively)[14]. These assign a base type to each noun and verb in the database, such as 'nouns denoting actions', 'nouns denoting feelings', 'verbs of thinking', 'verbs of telling', and so on. The lists include 61,686 nouns and 9719 verbs, ensuring that a large proportion of the verbs and nouns encountered will be assigned to a category. For example, *user* belongs to category 18, meaning it is a 'noun denoting people'; *caterpillar* to category 5, 'animal'; *smell* to category 39, 'verbs of perception'.

However, several lexical items are assigned to more than one category, ranging from just two, for example *risk*, which is both a 'noun denoting act or actions' and a 'noun denoting attributes of people and objects', to almost all available ones, as for the verb *run*, which belongs to twelve categories out of a possible fifteen. Leaving aside the issue of whether these category assignments do correspond to our intuitions, such a proliferation of categories raises the question of their usefulness. In these cases, in the current implementation of DAPPER, a separate contextual predicate is created for each category. If so many words can belong to most categories, this might make it harder for the classifier to discriminate among terms, as differences between contexts would be minimised. It might be advisable to assume a cut-off point for these category assignments, for example list only the first one or two.

The verb subcategorisation frame information is only used for the preposition component. It is derived from a list originally compiled using the dictionary distributed with the Alvey Natural Language Tools[15]. Unlike the WordNet data, the verbs included are assigned to just one category; however, the list includes only 688 verbs, so it is possible that as these represent only a small fraction of the lexical items encountered, this information will be of limited usefulness for the classifier as broad generalisations cannot be drawn.

Finally, the determiner component includes information on whether the head noun is classified as count or mass noun, or 'either'. This information is derived from the CuvPlus English dictionary (see e.g. Mitton 1992), based originally on the Oxford Advanced Learner's Dictionary of Current English (Hornby 1974), and includes just over 72,000 nouns, which should ensure a thorough coverage of the data.

[14] http://wordnet.princeton.edu/man/lexnames.5WN
[15] http://www.cl.cam.ac.uk/research/nl/anlt.html

3.2.4 Some examples

An example will be now worked through to show more clearly the steps carried out by the scripts and how the information is extracted, using the sample sentence below; the parser output for it is shown in Figure 3.2.

(3) The Cumbrian birds arrived in the late 1990s from a breed-and-release programme of endangered birds.

```
# this file was generated by the following command(s):
#   /home/scratch/candc-1.00/bin/candc --models /home/scratch/models/

(ncmod _ birds_2 Cumbrian_1)
(det birds_2 The_0)
(ncmod _ 1990s_7 late_6)
(det 1990s_7 the_5)
(dobj in_4 1990s_7)
(ncmod _ arrived_3 in_4)
(ncmod _ programme_11 breed-and-release_10)
(det programme_11 a_9)
(ncmod _ birds_14 endangered_13)
(dobj of_12 birds_14)
(ncmod _ programme_11 of_12)
(dobj from_8 programme_11)
(ncmod _ arrived_3 from_8)
(ncsubj arrived_3 birds_2 _)
<c> The|the|DT|I-NP|O|NP[nb]/N Cumbrian|cumbrian|JJ|I-NP|O|N/N
birds|bird|NNS|I-NP|O|N arrived|arrive|VBD|I-VP|O|S[dcl]\NP
in|in|IN|I-PP|O|((S\NP)\(S\NP))/NP the|the|DT|I-NP|I-DAT|NP[nb]/N
late|late|JJ|I-NP|I-DAT|N/N 1990s|1990s|NNS|I-NP|I-DAT|N
from|from|IN|I-PP|I-DAT|((S\NP)\(S\NP))/NP a|a|DT|I-NP|I-DAT|NP[nb]/N
breed-and-release|breed-and-release|JJ|I-NP|I-DAT|N/N
programme|programme|NN|I-NP|O|N of|of|IN|I-PP|O|(NP\NP)/NP
endangered|endangered|JJ|I-NP|O|N/N birds|bird|NNS|I-NP|O|N
.|.|.|O|O|.
```

Figure 3.2: Parser output for sample sentence

From the parsed and tagged output, the script creates feature vectors for the relevant items. These then undergo some further pre-processing to ensure uniformity, including turning class labels into lower-case, turning all instances of *an* into *a*, and removing whitespace and other extraneous formatting. Furthermore, the scripts create feature vectors for every lexical item encountered which is tagged as a preposition

or a determiner; only those for the class labels of interest are extracted. In Figure 3.3 we can see the final, classifier-ready feature vectors for the three prepositions in the sample sentence.

```
from 3left:DT 2left:JJ 1left:NNS 1right:DT 2right:JJ 3right:NN
mod_verb ModVerb:arrive WN_Class_ModV:38 WN_Class_ModV:41
Subcat_frame:4 ModGR_type:ncmod obj_noun ObjNoun:programme
ObjGR_type:dobj ModV_morethanone

in 3left:JJ 2left:NNS 1left:VBD 1right:DT 2right:JJ 3right:NNS
mod_verb ModVerb:arrive WN_Class_ModV:38 WN_Class_ModV:41
Subcat_frame:4 ModGR_type:ncmod obj_noun ObjNoun:1990s NE_ObjN:DAT
ObjGR_type:dobj ModV_morethanone

of 3left:DT 2left:JJ 1left:NN 1right:JJ 2right:NNS 3right:. mod_noun
ModNoun:programme ModGR_type:ncmod obj_noun ObjNoun:bird
WN_Class_ObjN:05 WN_Class_ObjN:06 WN_Class_ObjN:13 WN_Class_ObjN:18
WN_Class_ObjN:20 ObjGR_type:dobj
```

Figure 3.3: Preposition feature vectors for sample sentence

Let us take a closer look at the second one to gain a clearer picture of the various components of the feature vector. The first element is the class label, in this case ***in***. The next six contextual predicates are the POS tags of the three words to the left and right of the preposition, which are recovered from the POS-tagged sentence of the parser output. We then see, through the GRs, that the preposition modifies the word *arrived*, which is a verb belonging to two WordNet classes, 38 ('verbs of motion') and 41 ('verbs of political and social activities and events' – as noted above, the class assignments may not always correspond to one's intuitions). It is also assigned to subcategorisation frame 4, 'intransitive'. The preposition is in a non-clausal modifier relation with the verb it modifies ('`ModGR_type`'), as can be seen from the GR output of the parser. The object information can also be recovered from the GRs; the object in this case is a noun, the date *1990s*, which is also a named entity, as we can see from the tagged sentence. The preposition and its object are in a direct object relation (*dobj*), as is almost always the case. As the object does not have a WordNet entry, there is no information about its category, unlike for example the feature vector representing ***of*** which follows it. Finally, the feature '`ModV_morethanone`' notes the fact that the verb modified by the target preposition is modified by at least one other PP, in this case ***from***.

```
the HeadNoun:bird WN_Class_N:05 WN_Class_N:06 WN_Class_N:13
WN_Class_N:18 WN_Class_N:20 NounType:count HeadNtag:NNS
ModbyAdj:Cumbrian AdjTag:JJ 1right:JJ 2right:NNS 3right:VBD

the HeadNoun:1990s HeadNtag:NNS NEnoun:I-DAT ModbyAdj:late AdjTag:JJ
3left:NNS 2left:VBD 1left:IN 1right:JJ 2right:NNS 3right:IN
ObjofPrep ObjofPrep:in

a HeadNoun:programme NounType:count HeadNtag:NN
ModbyAdj:breed-and-release AdjTag:JJ 3left:JJ 2left:NNS 1left:IN
1right:JJ 2right:NN 3right:IN ObjofPrep ObjofPrep:from ModbyPrep
ModbyPrep:of

NULL HeadNoun:bird WN_Class_N:05 WN_Class_N:06 WN_Class_N:13
WN_Class_N:18 WN_Class_N:20 NounType:count HeadNtag:NNS
ModbyAdj:endangered AdjTag:JJ 2left:DT 1left:JJ 1right:VBD 2right:IN
3right:DT
```

Figure 3.4: Determiner feature vectors for sample sentence

The determiner feature vectors, seen in Figure 3.4, have a similar structure. Where a determiner is present, the head noun linked to it is easily derived from the GRs labelled 'det'. However, there is also a large number of bare NPs, i.e. without a determiner, such as mass nouns and plurals. These are retrieved by isolating the NPs in the parsed sentence; these phrases are also used as the local context for other features, as noted above. The second instance of ***the***, for example, tells us that the head noun is tagged as a plural common noun ('NNS'); the tagger does not recognise it as a proper noun, but the NER does, and correctly labels it as a 'date' named entity ('`NEnoun:I-DAT`'). The feature vector then records that the head noun is modified by the adjective *late*, which here is in its base form, and gives the POS tags of the three-word window surrounding the determiner. Finally, we see that it is the object of a preposition ('`ObjofPrep`'), in this case the preposition *in*. In the following feature vector, for *a*, we can also see that where present, information about noun type is recorded – in this case, that *programme* is a count noun – and that it is modified by a preposition as well as being the object of one. In the final example, there is an instance of a null determiner. Both here and in the first feature vector, the head noun *bird* is also assigned to the relevant WordNet classes, which are 5 ('animals'), 6 ('man-made objects'), 13 ('food'), 18 ('person'), and 20 ('plant') – exemplifying the danger of having a proliferation of contextual predicates related to this dimension,

not all of which are uncontroversial.

3.3 The datasets

This section gives a brief overview of the composition of the datasets obtained through the processes outlined above, and compares their composition to that of others found in the literature.

3.3.1 Prepositions

As mentioned in Chapter 1, this work is based around the nine prepositions ***at, by, for, from, in, of, on, to***, and ***with***, as these are the most frequent in the data. They also display the most variability in use and consequently cause the most problems for learners, so it seems reasonable to focus on this set. The basic training set consists of nearly nine million feature vectors, or instances: 8,989,359. This figure is not distributed uniformly across all nine classes, however: ***of*** is by far the most frequent, while others such as ***at*** or ***from*** are comparatively less frequent. The complete distribution figures for each item within the training set are show in Table 3.3. An important point to be noted is that instances of ***to*** include both its use as a preposition and as an infinitival marker. These receive the same tag from the POS tagger ('TO' as opposed to the general preposition tag 'IN') and it was decided not to filter out cases of infinitival marking, as misuse of the particle in these contexts is also often found in L2 writing.

Although these figures cannot be compared directly with those found in related work, because of the different types of data used as a source, it is noted that in Gamon et al. (2008) thirteen prepositions are dealt with, chosen on the basis of their observed frequency in an error corpus; this set includes all nine used in the present work. The figure for the size of the training set, which is extracted from a set of over 1.5 million sentences, is not given. The set of prepositions used by Lee and Knutsson (2008) is most similar to the present one, as it includes the same nine prepositions considered in the present work plus a tenth one, ***as***. Their training set consists of 10 million sentences, but a figure for the number of instances used is not given. Chodorow et al. (2007) train on 7 million instances, covering 34 different prepositions. The much greater number of possible classes, together with the smaller overall size of the training set, implies that for each preposition, a smaller amount of data will be available. These considerations must be borne in mind in carrying out any comparisons between these approaches and the one presented here.

	Training size	Percentage of total
at	424,539	4.72%
by	421,430	4.69%
for	720,369	8.01%
from	347,105	3.86%
in	1,589,718	17.68%
of	2,501,327	27.83%
on	587,871	6.54%
to	1,855,304	20.64%
with	541,696	6.03%

Table 3.3: Distribution of prepositions in training

3.3.2 Determiners

For the determiner task, the analysis is restricted to just three cases, ***a/an***, ***the***, and the ***null*** case. The basic training set consists of 4,043,925 instances. As for the prepositions, here, too, the distribution across the three classes is not uniform. The null case is by far the most frequent, as shown in Table 3.4; these relative frequencies are roughly replicated throughout the literature.

	Training size	Percentage of total
a	388,476	9.61%
the	1,180,435	29.19%
null	2,475,014	61.20%

Table 3.4: Distribution of determiners in training

As noted in Chapter 2, the vast majority of related work on determiner selection focuses on these three classes only. Knight and Chander (1994) is a notable exception, dealing with the two-way choice between definite and indefinite determiner; their source consists of 400,000 NPs. Minnen et al. (2000) use even fewer, just over 300,000, while the figure quoted by Lee (2004b) is 260,000. The work in Han et al. (2006), on the other hand, is on a different scale, as there 6 million NPs are used, while Turner and Charniak (2007), who use a language model, train on up to 20 million words.

In this chapter, the theoretical assumptions underpinning these investigations were described, along with the methodology used in preparing suitable data. In designing the feature sets to be used in creating the feature vectors, linguistic intuitions have been followed. In the next chapter, the results obtained by DAPPER on the classification tasks will be presented; the contribution of the individual features will be quantified more precisely in Chapter 5.

Chapter 4

The L1 model: results and discussion

This chapter presents and discusses the results obtained by DAPPER on L1 prepositions and determiners on various tasks. Section 4.1 presents results from the first, basic experiments. The remaining sections regard variations on the basic training model: Section 4.2 discusses results obtained by modifying some of the parameters available to the classifier, while Section 4.3 examines performance on individual prepositions and determiners.

4.1 Reference results

In this section, the 'basic' task giving reference results against which to compare all other figures reported is described. For both prepositions and determiners, these refer to data tested on a model resulting from 600 iterations in training, and where no features have been omitted, including those occurring only once. The figure of 600 iterations was found to offer a good trade-off between performance and training time, and so represents a good starting point from which to base further observations.

4.1.1 Prepositions

As described in Section 3.3, the training set for prepositions consists of nearly 9 million instances. The feature space is very large and diverse: there are 321,377 distinct contextual predicates occurring in the training data, with frequencies varying from just 1 to over 8 million. The classifier does not output a figure for the total number of contextual predicates present, which has been calculated at over 177 million (177,379,700). On average, then, each contextual predicate occurs about 550 times; in

Author	Accuracy
Baseline	26.94%
Gamon et al. 08	64.93%
Tetreault and Chodorow 08a,b	79.00%
Lee and Knutsson 08	70.20%
Training data	70.58%
Human task	88.60%
DAPPER	**70.06%**

Table 4.1: Classifier performance on prepositions - L1 data

fact there is a very large number of *hapax legomena*, as well as a small set of features which occur several million times[1].

The model developed with this training configuration is assessed by a test set consisting of a section of the BNC not used in training (all texts in the 'J' section), which comprises 536,193 instances. These instances were created using the same procedure outlined in Section 3.2.3.2 for the training feature vectors, meaning that all occurrences of the nine target prepositions in this section of the BNC are each represented by a feature vector with the original preposition as the class label. For each such feature vector, the class label is removed and the classifier's task is to assign a class label – one of the nine possible prepositions – to the feature vector. If its choice matches the preposition found in the original text, it is considered correct. An analogous procedure is used for the determiner task, the results of which are discussed in Section 4.1.2.

The success of the classifier on this task is measured by considering the number of times it has correctly assigned a class label to an instance; this is referred to as accuracy. On the test data, accuracy is **70.06%**. This could be thought of as analogous to a precision score as it gives the measure of the proportion of relevant or correct responses over the total of responses given, out of a possible total of 100%. Performance in terms of recall and precision will be discussed in greater detail in Section 4.3, where these measures will be analysed with reference to individual prepositions.

In fact, accuracy may not be the most informative measure for assessing DAPPER's performance. Because there is an imbalance within the data, with some classes more heavily represented than others, this measure will be skewed towards the results

[1]There are in fact only 15 contextual predicates which occur 2 million times or more; another 28, mostly referring to WordNet classes, have a frequency between 1 and 2 million. The top five are, in order: *objectGR:dobj, object_is_noun, modifiedGR:ncmod, modified_is_verb, modified_is_noun.* The other contextual predicates in this group refer to the POS context, to other GRs, or to the verb being transitive or modified by more than one PP.

obtained on the larger classes. If these are high, and those of the smaller classes are lower, we will have a distorted representation of the system's overall success. Precision and recall, discussed in Section 4.3, will give a clearer picture of variations in performance across the classes. By looking at these scores, we can assess the extent to which data imbalances may be affecting the data and introducing noise. The accuracy figure reported here quantifies the system's error rate; precision and recall offer the means to understand the possible causes of such errors.

To put this figure of 70.06% in context, we can refer to Table 4.1. The baseline, which refers to always choosing ***of***, the most frequent class, is significantly outperformed. This is not surprising: as there are nine different classes to choose from, it is unlikely that this type of guessing would be successful very often. The figure from a task run with human testers is also included. This was designed to estimate what a likely upper bound might be for this task, given the challenges posed by preposition selection (as discussed in Section 1.1). Two native speakers of British English were asked to complete a task analogous to that submitted to the classifier, namely selecting one of the nine target prepositions to complete a set of 841 contexts from which they had been removed. Their accuracy, measured in terms of agreement with the original text, averaged 88.6%, proving that this task is hard for humans, too. As further evidence of the variability present in preposition selection, it is also noted that agreement between the two subjects is around 87%[2].

The figure for training data is also included – here and elsewhere in the discussion. It is useful to compare accuracy on training and testing data, to establish whether the model developed suffers from overfitting on the training data. The table shows that there is little difference between accuracy on training and testing data, which means that the model can perform more or less equally well both on novel and on seen data.

As mentioned above, other work in the literature achieves similar accuracy. It is not possible to draw direct comparisons between Dapper and these models, however, because of several differences present in the research setup. As noted in Section 3.3, each of the models is trained to recognise a different number of prepositions. Furthermore, each group uses a different approach in deciding what contextual predicates to extract (cf. Section 3.1.4.1) so that there is not much overlap in the feature sets, and the training data is taken from different types of sources (British vs. American English, BNC vs. news reports, and so on). While Tetreault and Chodorow (2008a) use

[2]Cohen's kappa for Subject 1-text = 0.867, Subject 2-text = 0.860; intersubject agreement is kappa=0.842. All three scores are considered 'very good'.

a maximum entropy classifier, this is not the case for the Gamon et al. (2008) work, which uses a decision tree combined with a language model, or Lee and Knutsson (2008), who use memory-based learning.

It should be also noted that Gamon et al. report more than one figure in their results, as there are two components to their model: one determining whether a preposition is needed, and the other deciding what the preposition should be. The figure referred to here refers to the latter task, as it is the most similar to the one being evaluated. Furthermore, the figure given for Tetreault and Chodorow refers to a model which incorporates several filters to refine results, something which DAPPER does not currently make use of. A more basic version of their model achieves around 69% accuracy (Chodorow et al. 2007).

Given the methodological differences described so far, the figures in Table 4.1 cannot be used to draw any firm conclusions as to the superiority or otherwise of one particular system. Rather, they illustrate that, while such approaches to preposition usage acquisition are viable, the task is challenging for humans and NLP systems alike, and the optimal way of addressing it is still being developed.

4.1.2 Determiners

The training set for the determiner task, as noted previously (Section 3.3), consists of 4,043,925 instances. Again, the feature space is very large. 185,959 different contextual predicates are found in the training data, occurring overall a total of 52,596,314 times, with the most frequent having a frequency of 1 to 2 million[3]. Testing is on data extracted from a section of the BNC not seen in training (Section 'J'), comprising 305,264 instances. From one perspective, this classification task might be expected to receive a higher accuracy than the preposition one because the classifier must only choose from among three rather than nine possible classes. On the other hand, as seen, determiner choice is less fixed than preposition choice: while not all lexical items can appear with all prepositions, for example, the fact that determiner occurrence depends less on short syntactic and lexical sequences and more on discourse features means that it is possible that the same head nouns will appear in the training

[3]Nine contextual predicates occur between 1 and 2 million times. The top five are, in order of frequency: *head noun tag:NN*, *noun is object of preposition*, *nountype:count*, *nountype:either*, *POS tag to the left is IN*. The remaining are related to WordNet. A further 15 contextual predicates occur between 500,000 and 1 million times. These include more WordNet information, head noun tags NNS and NNP, POS tag context, and another three preposition-related contextual predicates: modified by preposition, object of preposition 'of', and modified by preposition 'of'. The latter two provide further evidence of this preposition's high frequency in the language.

Author	Accuracy
Baseline	59.83%
Han et al. 06	83.00%
Gamon et al. 08	86.07%
Turner and Charniak 07	86.74%
Training data	93.26%
DAPPER	**92.15%**

Table 4.2: Classifier performance on determiner task - L1 data

data in conjunction with two or all three classes, potentially making it harder for DAPPER to select the correct label.

Accuracy achieved on this task is **92.15%**, as shown in Table 4.2, where this figure is compared to other recent results in the literature. Again, the baseline refers to the most frequent class, which is the ***null*** case. Although the baseline is already comparatively high, this result represents an absolute improvement of over 30% on it, proving the viability of the approach. Indeed, the difference with the training data figure is minimal in this task also, which is further evidence of the robustness of the model developed.

With regard to comparisons with the other approaches mentioned, similar caveats to those mentioned in the previous section apply in establishing the extent to which the results can be compared to each other. Although, as already observed (Chapter 2), in work on determiners there is much less variation in terms of the kinds of contextual predicates chosen, and all recent work considers the same three classes, the data sources for training and testing remain different. It is hypothesised that the good performance of DAPPER may be due to the contribution of contextual predicates that give a fuller picture of the NP's context, such as NEs and PP and adjectival modification, but this cannot be ascertained without a more direct comparison between the various approaches.

4.2 Classifier parameters: variations and results

This section discusses variations in the classifier training procedure, to assess whether the optimal setup for the learning procedure has been selected.

Nr. of training instances	Training data acc.	Test data acc.
2,471,191	71.01%	69.21%
4,860,314	71.06%	69.62%
6,495,667	70.97%	69.84%
8,989,359	70.58%	70.06%

Table 4.3: Effect of training size on accuracy

4.2.1 Prepositions

4.2.1.1 Training set size

As noted in Chapter 3, to train the model for the preposition task a very large dataset is used, comprising almost 9 million instances. Of course, working with such a large training set can be computationally expensive, so if reliable results could be achieved with a smaller dataset, this would be advantageous. On the other hand, it is generally acknowledged that in training such classifiers, having more data is a better guarantee of obtaining representative results with fuller coverage. To address this issue in the context of the preposition task, the classifier is trained on datasets of differing sizes, as reported in Table 4.3.

The results of the four tasks using different amounts of training data do not display much variation. However, further investigations are necessary to establish the point at which the classifier begins to tail off: it may be that a similar accuracy may be achieved with a much smaller dataset. It is possible that there is a natural 'ceiling' for this task which cannot be overcome even by a major increase in the amount of training data used, or that the BNC is not the best source of data after all.

4.2.1.2 Iterations

To establish the number of iterations used in all the experiments, various tasks were attempted with this figure varying from 400 to 2400. As little difference was found in the results, however, 600 was chosen as a good compromise between training time and accuracy. Table 4.4 illustrates this point further by showing the results achieved by the models with the various increases in iterations.

4.2.1.3 Feature pruning

An important option made available by the classifier is that of specifying the 'cutoff' parameter, which allows the minimum frequency cutoff point for features to be specified. This is set at 1 by default, and, as reported in Section 4.1, this is the setting used

Nr. of iterations	Training data accuracy	Test data accuracy
400	70.39%	69.90%
600	70.58%	70.06%
800	70.66%	70.10%
1000	70.70%	70.12%
1600	70.75%	70.17%
2400	70.78%	70.20%

Table 4.4: Effect of number of iterations on preposition accuracy

Cutoff point	Training data accuracy	Test data accuracy
no cutoff	70.58%	70.06%
2	70.32%	70.03%
11	69.52%	69.79%

Table 4.5: Effect of cutoff threshold on accuracy

in the basic, 'reference' experiments. However, it is common practice to disregard features occurring below a given frequency threshold, as very infrequent features may introduce unwanted noise and potentially obscure relations between instances that should be classified together. Feature pruning also leads to a reduction of the size of the feature set, which can make training and testing faster. These claims are tested by training two models, one where the cutoff point is 2 – i.e. features occurring only once are removed – and one where the cutoff is 11, removing all features occurring ten times or less. Results are reported in Table 4.5.

Strikingly, the assumption that removing low-frequency features would assist performance is not borne out by these results. Only minimal variation is found, with the lowest results associated with the greatest cutoff point. Perhaps it must be assumed that even features which occur only 10 times over a set of several thousand make a meaningful contribution to the overall model[4], although this would not explain the decrease in accuracy for the cutoff 2 set.

In fact, it is not necessarily the case that simply excluding low frequency or even singleton features does improve performance. In Curran and Clark (2003a), for example, the assumption that such features are "unreliable or uninformative" is tested by comparing tagging results obtained by this method and by using a Gaussian prior instead. The authors try a variety of cutoff values, but find that the best results are

[4]A quick and by no means complete look at the features which fall into this group proves inconclusive. They almost all involve lexical items occurring as heads or objects of the preposition. While some are fairly common words, which could conceivably have quite strong collocational patterns (e.g. *admittance*, *limelight*, *rudeness*, *clench*), others are decidedly uncommon, either proper names or generally rare words (e.g. *kaolin*, *Grunwick*, *Medvedev*, *organochlorine*).

Cutoff point	Distinct contextual predicates	Acc.
no cutoff	321,377	70.06%
2	125,814	70.03%
11	34,876	69.79%

Table 4.6: Effect of cutoff threshold on prepositions feature set

had when no cutoff is imposed. Their conclusion is that, rather than feature pruning, Gaussian smoothing (included in the maximum entropy algorithm used here) helps achieve optimal performance, and suggest that even low frequency features have a contribution to make to results. These findings are in line with those reported in Daelemans et al. (1999), where results from a variety of NLP tasks lead the authors to conclude that even "exceptional events can be beneficial for accurate generalisation" (Daelemans et al. 1999:31). In conclusion, since even single or rare features contribute to the weighting of other features, it is important that they be retained.

A related observation regards the size of the feature space. Removing unique features leads to a decrease in the number of contextual predicates overall, as can be seen in Table 4.6.

It is immediately evident that infrequent features make up a large part of the feature space, as the figure for the cutoff point 11 shows: this set is nearly 90% smaller than the full one, which makes for smaller files and much faster training and testing. Is there a linguistic explanation for these figures? Of course, language is a creative and open system, which means that any number of words can appear as heads and objects of prepositions in combinations which may have never been heard before (and indeed may never be heard again). This accounts largely for the *hapax legomena* features. Furthermore, errors on the part of the POS tagger or stemmer cannot be ruled out, for example labeling the same item sometimes as a noun and somtimes as a verb, which would introduce extraneous features to the set. Overall, despite the removal of such a large number of features, the contextual patterns that remain are sufficiently robust to account for the majority of cases encountered, lending support to the notion that these patterns can indeed be abstracted from the lexical items involved.

4.2.2 Determiners

4.2.2.1 Training size

The size of the determiner set used for training is smaller than that of the prepositions; however, especially in the light of the amount of training data used by several others

Nr. of training instances	Training data acc.	Test data acc.
1,206,787	92.63%	91.32%
2,338,552	93.40%	92.61%
3,118,009	93.37%	92.69%
4,043,925	93.26%	92.15%

Table 4.7: Effect of training size on determiner accuracy

Nr. of iterations	Training data accuracy	Test data accuracy
400	93.19%	92.09%
600	93.26%	92.15%
800	93.30%	92.18%
1000	93.32%	92.21%
1600	93.35%	92.24%
2400	93.36%	92.26%

Table 4.8: Effect of number of iterations on determiner accuracy

in the literature, which varies from a few hundred thousand to several million (cf. Section 3.3), it is informative to compare performance on different sizes of training sets for this task, too, and examine the extent to which results are affected by increases in training data. This is summarised in Table 4.7.

Here, too, no great variation among the four results is found, especially with reference to the training data. Furthermore, the same observation applies here as in the analogous discussion regarding the prepositions task in Section 4.2.1.1, namely that accuracy could have begun to plateau at a much smaller number of training instances.

4.2.2.2 Iterations

As for the prepositions task, iteration variation experiments were repeated for this POS. Again, little difference was found in the results, and the figure was set at 600 for all experiments reported. Table 4.8 presents the results achieved by the models with the various iteration amounts.

4.2.2.3 Feature pruning

In Section 4.2.1, it was observed that imposing a cutoff point for low frequency features brings only a minor decrease in accuracy. It is informative to see whether this also occurs for determiners, which would confirm the hypothesis that the algorithm itself can account for the presence of these features, or if it is just a peculiarity of

Cutoff point	Training data acc.	Test data acc.
no cutoff	93.26%	92.15%
2	93.18%	92.14%
11	92.96%	92.06%

Table 4.9: Effect of cutoff threshold on determiner accuracy

Cutoff point	Distinct cont. predicates	Acc.
no cutoff	185,959	92.15%
2	86,850	92.14%
11	22,826	92.06%

Table 4.10: Effect of cutoff threshold on determiner feature set

prepositions. The task is therefore repeated for this POS, once again setting the cutoff point at 1 (full feature set), 2, and 11. Results appear in Table 4.9.

The outcome is analogous to that of the preposition task, indeed here the difference in the test data results is even less marked. This suggests that, if greater processing speed were required, singleton features could be sacrificed without too great a negative effect. A quick scan of the feature list shows that the features removed are almost all related to the lemma of the head noun, which is unsurprising, as there are of course millions of different nouns in the BNC, any number of which could occur only once with the three determiner classes considered. Their removal should not affect accuracy too much, then, because the classifier would still have all the more general contextual information to rely on to build its model – and indeed this is what is aimed for, a model that is more pattern- than lemma-dependent.

It is useful to assess the effect of feature pruning on the feature space, as presented in Table 4.10. Once again, at the cutoff point of 11, the feature set is barely more than 10% of the full one. However, the removal of these low-frequency items does not have a positive effect on the results, so it might prove more advisable for the moment to retain all features in training.

4.3 Individual items: results and discussion

A full appraisal of Dapper's performance cannot exclude the analysis of its accuracy on the individual preposition and determiner classes it is concerned with; the system would be of little use if its high scores derived only from good accuracy on one or two classes to the detriment of all others. A related issue of interest is whether, should imbalances in accuracy be found, these are due to flaws in the models or to inherent

properties of the items in question: for example, are certain prepositions perhaps more 'learnable' than others? In this section, an analysis of DAPPER's results broken down by classes is presented.

4.3.1 Prepositions

In discussing results pertaining to individual preposition classes, it is possible to calculate precision and recall. Recall here refers to the proportion of correct answers given out of all possible correct answers to be had:

$$Recall = \frac{\#\ of\ correctly\ labelled\ instances}{\#\ of\ instances\ to\ be\ labelled}$$

In other words, if there are 100 instances of the preposition ***at*** to be identified, and DAPPER correctly assigns the label ***at*** to 75 of them, recall will be 75%. Precision refers to the proportion of correct or appropriate answers for a given class, out of all the answers given for that class:

$$Precision = \frac{\#\ of\ correctly\ labelled\ instances}{\#\ of\ times\ class\ label\ assigned}$$

With specific reference to the preposition task, for example, if DAPPER has labelled 50 instances as belonging to the class ***at***, but only 25 of those are actually instances of ***at***, its precision will be 50%. These measures are important because the aim is to achieve a high score for both: a system which had a high recall, but at the expense of poor precision (e.g. by choosing certain classes far more than needed) would be of little use, especially in the context of a learner-oriented application, as discussed more fully below. A high recall score for this task means that DAPPER is able to correctly assign labels to a large number of instances; a high precision score indicates that the results are relatively free of noise. In the evaluation, the F-score is also considered. This is the harmonic mean of precision and recall ($F = \frac{2(precision \times recall)}{(precision + recall)}$), which gives a better overall assessment of performance.

In Table 4.11, precision and recall figures for each preposition (test data) are reported, as well as absolute and relative frequencies in the training data. It is important to include this latter figure, as it can help establish if there is any correlation between how many instances of a given class are present in the training set, and DAPPER's performance on that particular class in testing. In the table, the prepositions are arranged in descending order of frequency, and we can see immediately that this correlation is not straightforward. From the two most frequent cases, ***of*** and ***to***, it would seem evident that observing a high number of instances in training leads to

	Proportion of training data	Precision	Recall	F-score
of	27.83% (2,501,327)	74.28%	90.47%	81.58%
to	20.64% (1,855,304)	85.99%	81.73%	83.81%
in	17.68% (1,589,718)	60.15%	67.60%	63.66%
for	8.01% (720,369)	55.47%	43.78%	48.94%
on	6.54% (587,871)	58.52%	45.81%	51.39%
with	6.03% (541,696)	58.13%	46.33%	51.39%
at	4.72% (424,539)	57.44%	52.12%	54.65%
by	4.69% (421,430)	63.83%	56.51%	59.95%
from	3.86% (347,105)	59.20%	32.07%	41.60%
macroaverage		63.67%	57.38%	59.68%

Table 4.11: Individual prepositions results - test data

good results for both precision and recall: in other words, it allows the model to build a reliable representation of these prepositions' contexts. Other cases, however, are not so clear-cut. Excluding the three classes which have more than 1 million instances, for example, the highest scores are found to be for ***by***, which is the second-smallest set. Conversely, one of the lowest scores is found with ***for***, which is among the more frequent prepositions. Nor is having over 1 million instances a guarantee of high recall: in the case of ***in***, while its scores are indeed rather higher than the other less frequent prepositions, they are not so close to those for ***to*** as the similar frequency might lead one to expect. However, as previously discussed, these considerations of the amounts of data available for each preposition would need to be further supported by observations of the points at which performance has plateaued for each of them.

The absence of a definite relation between frequency in training and recall figures means one must look further for an explanation to the fluctuations in the latter. It can be argued that there is a 'learnability cline': different prepositions' contexts may be more or less uniquely identifiable, or the prepositions may have more or fewer senses, leading to less confusion for the classifier. One simple way of verifying the likelihood of the latter case is to look at the number of senses assigned to each preposition by a resource such as the Oxford English Dictionary (OED). However, this does not yield any informative correlations: the preposition with the most senses is ***of*** (16), and that with the fewest is ***from*** (1), precisely the reverse of what would be expected were the relation true[5]. The reasons for the wide-ranging differences must therefore lie elsewhere; for example, it is more likely that different prepositions may be found as labels for very similar feature vectors.

[5]Although of course there is no guarantee that the OED's partitioning of senses is the one most relevant to the needs of this task.

Perhaps not all the prepositions have contexts whose lexical and grammatical characteristics allow unique and unambiguous label assignment. This is intuitively appealing; after all, while a sentence containing a passive form such as *she was bitten — a snake* can only plausibly be completed by the insertion of ***by*** (or perhaps, less commonly, ***near***), in *give it to the person — the office* we could use ***at***, ***from***, ***in***, or ***for***. The difference between these two examples can be thought of as expressing the difference between functional or grammatical use and semantic use[6].

Gamon et al. (2008) report some similar observations on differences in performance across prepositions, and also give a breakdown of results according to individual items. While there is broad agreement with their work as regards the best-performing prepositions (***of, in, to***), the present research's results differ noticeably for the other cases, which could be due either to difference in training data or in the methods used. In Table 1 of Tetreault and Chodorow (2008a), the authors compare the F-measures obtained for individual prepositions with those achieved by Gamon et al. (2008); this table is reproduced here (Table 4.12) together with DAPPER's F-scores for a three-way comparison, although it must be noted that while the latter scores refer to BNC data, those in the other two papers refer to data from Reuters News and Microsoft Encarta.

Because of the differences in the data used noted above, it is impossible to compare DAPPER's score directly to the other two, although it would suggest that a sophisticated and feature-rich model may have some advantages over a simpler approach. The scores obtained by Tetreault and Chodorow (2008a) would seem to indicate that the differences in preposition learnability may not be as great as hypothesised; they also use a variety of postprocessing filters to improve results, which may be crucial in ensuring a more uniform performance.

So far the discussion has focused mostly on issues of recall, but of course it is important to analyse precision figures, too. Indeed, arguably precision is to be favoured over recall for this task. Naturally one would aim for high figures for both measures, but if one has to be privileged over the other, it makes sense for it to be precision. This is because, in the context of error correction, it is preferable to miss out a few cases of misused prepositions than to raise false alarms and impose an unnecessary correction on error-free text, undermining the learners' confidence in their ability.

[6]This distinction is not very clear-cut (cf. discussion in Keizer (2004)), but the two categories can be thought of as referring to uses where a particular preposition is required by the lexical item to introduce its complement (functional use, e.g. *reliance on her father*) and those where the preposition carries some core locative/spatial/temporal meaning and introduces a modifier (semantic use, e.g. *the book on the table*). A more in-depth discussion of this issue is found in Tseng (2000:15-35).

	Gamon et al. 2008	Tetreault&Chodorow 2008b	Dapper
of	75.9%	90.6%	81.6%
to	62.7%	77.5%	83.8%
in	59.2%	84.5%	63.7%
for	40.5%	69.8%	48.9%
on	32.2%	75.1%	51.4%
with	36.1%	67.5%	51.4%
at	37.2%	68.5%	54.6%
by	50.2%	74.7%	59.9%
from	52.8%	59.1%	41.6%
macroaverage	49.64%	74.14%	59.7%

Table 4.12: Individual prepositions results – F-score comparison (data partly reproduced from Tetreault and Chodorow (2008a))

Overall, the precision scores are higher than the corresponding recall figures, which is positive, although they could be improved further. The one striking difference is for ***of***, where precision – while still high – is much lower than recall.

The reason for this may yet lie in the amount of data seen in training for this preposition. It was shown that it is possible that high recall is achieved not just because of more easily identifiable occurrence contexts, but also because the overwhelming frequency of this preposition is likely to make it the classifier's default choice when it is not able to establish what label should be assigned. Naturally, if ***of*** is relatively more frequent in the data, its instances will also be more frequent among the cases where this 'default' choice is made, and this will contribute to its better recall. But this would also contribute to lower precision, as it would be selected, again by default, in many more cases where it is not appropriate. There are two ways by which to test the validity of this claim, i.e. that high frequency of one item is skewing Dapper's performance: the analysis of a confusion matrix for the classifier's errors, and an investigation of its performance on a training set compiled to have equal amounts of instances for each preposition.

Imbalanced data sets, such as this one, tend to bias the classifier towards the majority class; this issue is widely discussed in the machine learning literature[7], where two main approaches are evident: modifying the composition of the data set, or introducing algorithmic modifications to the classifier without modifying the distribution of the data (see e.g. Maloof 2003; Japkowicz 2000; Chawla et al. 2004). Addressing the problem at the data level by resampling is more easily implemented in the context

[7]Albeit with the focus primarily on cases where the disparity in class size is of a greater magnitude than in the present case, and where the task is a binary classification one.

Nr. of training instances	Training data acc.	Test data acc.
2,471,191	71.01%	69.21%
reference - 8,989,359	70.58%	70.06%
UNDERSAMPLED - 3,123,936	64.55%	64.73%
OVERSAMPLED - 13,500,00	65.12%	66.05%

Table 4.13: Prepositions results - balanced training sets

of this task. Resampling can consist of either under- or over-sampling the data. In the first case, the size of the larger class is reduced so as to be more similar to the smaller class (see also Kubat and Matwin 1997). In the latter case, examples of the smaller class are duplicated so as to match the size of the larger class (see also Ling and Li 1998).

These approaches are found to work well in some contexts, though it is likely that their success is in part dependent on the characteristics of the data and of the task. In particular, it is less clear how well they can be adapted to the preposition classification problem, where there are more than two classes, and the imbalance across the classes is of varying size. Both under- and over-sampling also present some drawbacks. Undersampling may lead to the loss of potentially useful instances, while oversampling may lead to overfitting. Furthermore, both kinds of sampling change the distribution of the data, which may negatively affect the task to be performed.

To investigate possible effects of imbalanced data sets, both under- and over-sampling of the data is attempted. In the first case, just over 347,000 instances of each preposition are taken to create a training set. This is equal to the amount of instances of the preposition with the lowest frequency, ***from***, and yields a training set consisting overall of 3,123,936 instances, significantly smaller than the set of nearly 9 million which has been discussed so far. The evident disparity in size makes a direct comparison difficult, as size becomes a factor as well; a more realistic comparison would have equal amounts of each preposition, totalling 9 million instances, for the creation of which additional corpus resources would be required.

For over-sampling, instances from the less frequent prepositions are duplicated: a training set where each preposition is represented by 1.5 million instances is also created, giving a total of 13,500,000 training instances. Table 4.13 compares the results obtained from the resampled training sets to the reference results and to those obtained from a smaller, balanced training set, to facilitate comparisons with the undersampled set.

It is immediately evident that the results are lower than those of both imbalanced sets, despite the fact that the small equalised set has more training data available

Target prep	Confused with								
	at	**by**	**for**	**from**	**in**	**of**	**on**	**to**	**with**
at	xx	4.65%	10.82%	2.95%	36.83%	19.46%	9.17%	10.28%	5.85%
by	6.54%	xx	8.50%	2.58%	41.38%	19.44%	5.41%	10.04%	6.10%
for	8.19%	3.93%	xx	1.91%	25.67%	36.12%	5.60%	11.29%	7.28%
from	6.19%	4.14%	6.72%	xx	26.98%	26.74%	7.70%	16.45%	5.07%
in	7.16%	9.28%	10.68%	3.01%	xx	43.40%	10.92%	8.96%	6.59%
of	3.95%	2.00%	18.81%	3.36%	40.21%	xx	9.46%	14.77%	7.43%
on	5.49%	3.85%	8.66%	2.29%	32.88%	27.92%	xx	12.20%	6.71%
to	9.77%	3.82%	11.49%	3.71%	24.86%	27.95%	9.43%	xx	8.95%
with	3.66%	4.43%	12.06%	2.24%	28.08%	26.63%	6.81%	16.10%	xx

Table 4.14: Confusion matrix for prepositions

than the small imbalanced set. It was mentioned above that re-sampling affects the distribution of the data, and this has a negative effect for our task. The distribution in the imbalanced training data sets mirrors that of the language, which is that found in any test data, too. Therefore, despite the bias towards the larger classes observed in the imbalanced data set, it is more important that the classifier be trained to expect such distribution patterns rather than artificially imposed ones.

As for the confusion matrix, the relevant figures are summarised in Table 4.14 which reports, for each preposition, what the classifier's incorrect decision was, expressed as a percentage of overall incorrect decisions for that preposition. For example, instances of ***at*** were incorrectly labelled as ***by*** in 4.65% of cases, as ***for*** in 10.82% of cases, and so on. Analysis of these errors may go some way towards explaining whether they are related to frequency or have a more linguistically grounded motivation.

A frequency effect appears to be evident as, in almost every case, the three most frequent wrong choices are the three most frequent prepositions, ***to***, ***of***, and ***in***, although interestingly not in that order, ***in*** usually being the first choice despite being the least frequent of the three. This would suggest that the frequency effect is interacting with more subtle linguistic factors: the contexts for ***of*** and ***to*** are more clearly definable and perhaps more unique (e.g. more likely to modify adjectives, to link two nouns, to modify verbs of motion...), so the classifier is less likely to opt for those as default choices than for ***in***, which indeed has a relatively low precision score, compared to the other two. These observations are encouraging, as they point to a model which has acquired reliable linguistic generalisations as well as more basic frequency rules.

On the other hand, it is undeniable that frequency plays a role, as is made evident by the fact that the less frequent prepositions appear less often as the classifier's

choice. This makes it harder to come to any linguistic conclusions, as they have to be filtered through the frequency effects, though some facts do stand out. For example, one might expect the locative prepositions ***at***, ***in***, and ***on*** to be confused with each other more often than with others. While this is observed very clearly, reciprocally, for ***in*** and ***on***, the same does not happen for ***in*** and ***at***. Although ***in*** is very frequently given as the incorrect label for instances of ***at***, the reverse effect is not nearly as strong. Instead, ***at*** appears particularly frequently as the incorrect label for ***to***. This can be explained by the fact that both are spatial prepositions, one usually denoting a place as a stative point (*I was* ***at*** *the theatre yesterday*) and the other as a destination (*I went* ***to*** *the theatre yesterday*): it is easy to see how they could occur in very similar contexts.

Also within a spatial framework, further proof of strong contextual effects would be a clear confusion between antonymical prepositions such as ***to*** and ***from***, which are almost by definition expected to occur in identical contexts. Indeed, it is found that each preposition is the other's most frequent wrong choice, in other words, ***from*** is given most often as the incorrect answer for ***to***, and vice versa. This is evident even though ***from***, probably because of its low overall frequency, is only suggested as a label for instances of ***to*** less than 4% of the time: it is still a higher than average figure for this preposition, lending further support to the claim that frequency effects are interacting with contextual knowledge.

Less predictable, perhaps, is the undeniably strong connection between ***by*** and ***in***. Each stands out for the unusual frequency with which it is given as the incorrect label for the other – partly contradicting what was observed earlier regarding ***by***'s greater uniqueness and well-defined grammatical function. A quick overview of the contexts in which these confusions arise does not prove conclusive, as many kinds of different lexical items appear. One unifying thread seems to be the fact that most of the contextual patterns are of the form 'verb or deverbal noun - preposition - noun' (some examples: *indicated* ***by*** *the fact*, *produced* ***in*** *the state*, *succeed* ***in*** *becoming*, *interruption* ***by*** *war*), where either preposition could in principle be acceptable, depending on the intended meaning. This type of error, clearly driven by similarities in context, is unlikely to figure in errors made by learners, as the semantic function fulfilled by the two prepositions is rather different[8].

On the other hand, a more predictable semantic-driven confusion pattern, namely between ***by*** and ***with***, is not found: as both are often used to express instrumental

[8]Typical human learner error confusion patterns will be discussed in Chapter 6.

function, one might have expected overlap in their contexts. The fact that this does not occur suggests this overlap is not observed in the data.

By is not the only preposition that appears to be often confused with ***in***; the latter preposition is also shown in the table to be involved in confusion with ***of*** with very high frequency. The most likely explanation for this fact is that both prepositions occur very often between two nouns or NPs (*element* ***in*** *development*, *area* ***in*** *the Carpathians*, *steppe* ***of*** *Hungary*, *growth* ***of*** *trade*) where, as for the previous case, either preposition is grammatically possible, indeed often with little distinction in meaning. It is therefore not surprising that DAPPER cannot always choose a class label which is consistent with that found in the test data.

Another preposition to have a strong connection with ***of*** is ***for***, which is a somewhat unexpected finding with regard to linguistic intuitions – it might have been predicted that ***for*** would be confused more often with ***to*** as they are both used to indicate a goal or beneficiary[9]. An obvious similarity between the two is that, like in the previous case, they both occur between two nouns. However, several examples of phrases which have been incorrectly labelled by the classifier are listed here, as they might identify a more subtle shared feature of the two prepositions; the preposition in the phrase is the correct one, although it is immediately evident that the alternative in parenthesis is also grammatically acceptable:

- *basis* ***of*** *(for) stability*
- *mechanism* ***of*** *(for) distribution*
- *demand* ***of*** *(for) sector*
- *reservoir* ***for*** *(of) occupation*
- *condition* ***for*** *(of) development*
- *praise* ***for*** *(of) enterprise*

Several of these phrases involve deverbal nouns which can be thought of as having strong lexically-based cooccurrence patterns with both the prepositions, but would be used in different contexts.

A more predictable confusion pair is ***for*** and ***at***, where the erroneous label choices involve almost exclusively cases where the object of the preposition is a word such

[9]Indeed, ***for*** is given as the label for instances of ***to*** with slightly higher frequency than for some of the other cases, but not in a particularly exceptional way.

as *moment* or *time*: cf. ***at/for*** *the moment*, ***at/for*** *the time*, and so on. This distinction, which is hard to explain and is indeed generally not erroneous (since both phrases have a roughly equivalent meaning), would be expected also to prove confusing for human learners.

Interestingly, ***with*** appears to be almost never involved in any of these confusion pairs, despite comparatively low recall and not very high precision. The only two prepositions with which it shows a slightly higher frequency of confusion are ***to*** and ***for***. It is not easy to find conclusive evidence in the data, but in both cases, several examples seem to suggest strong lexical effects at play. These include phrases such as *link with, common with, come with* (incorrectly labelled as ***to***) and *provide with, leave with, be with* (labelled as ***for***). In all these examples, either preposition is grammatical, although in some cases the meaning changes; it can be hypothesised that the incorrect choice is dictated by a particularly strong correlation between the items and the other preposition. This is an example of contextual lexical effects being intensified by frequency, as both prepositions are more frequent in the training data than ***with***.

In this section, the performance of Dapper on individual prepositions was analysed, together with a confusion matrix showing the types of errors made. This analysis leads to the conclusion that while frequency of occurrence in the training data may bias some of the results, the model has acquired informative linguistic patterns, too, which can be found at the origin of several of its incorrect class assignments.

4.3.2 Determiners

As has been done for the prepositions, it is interesting to establish for the determiners also whether the high accuracy score is equally distributed across the three possible classes. The distribution of the training data across the three classes is unequal here as well, so it is important to ascertain the presence of any frequency effects. The results for each determiner class are shown in Table 4.15. Precision, recall, and F-scores are calculated in the same way as outlined for the prepositions in the previous section.

Here, even more than in the prepositions results, a strong correlation between the frequency of the class in the training data and results is evident. Precision and recall for the ***null*** case are high, as are those for ***the***; both classes being much more frequent in the data than ***a***. The latter determiner's lower 'learnability' appears not to

	Proportion of training data	Precision	Recall	F-score
a	9.61%	70.52%	53.50%	60.84%
the	29.19%	85.17%	91.51%	88.23%
null	61.20%	98.63%	98.79%	98.71%
macroaverage		84.77%	81.27%	82.59%

Table 4.15: Individual determiner results

be peculiar to this data, as it is also reported in Gamon et al. (2008) among others (F-scores in that work are: ***null*** 92.67%, ***the*** 90.72%, ***a*** 71.22%). However, it is unclear if this is due entirely to low frequency in training, or if there is something inherent in this determiner which makes it difficult to extract patterns of occurrence. After all, the precision for ***a*** is over 70%. Is this due to the model having some knowledge of when to suggest it as a class label, or is it to be explained by assuming that because of its relative low frequency, it is chosen less often overall and consequently there are fewer cases in which it is chosen in error? The low recall score would point to this latter explanation. On the other hand, as high precision is one of the goals of this work, the results obtained overall are encouraging.

The disparity among the three classes observed in the training data is a reflection of the distribution of determiners in the English language.The previous section discussed the advantages and disadvantages of having a balanced training set for the prepositions. Similar considerations apply here. Perhaps, if the imbalance were addressed, then the model would more confidently learn contexts of use for ***a***, too, which would be desirable in view of using this information for error correction. On the other hand, this would create a distorted representation of the composition of English, which is not what is wanted in a statistical model of language. Undersampling is used in the case of the determiners to assess the effects of rebalancing the training data.

Of course, because of the relatively small size of the ***a*** class, this will result in a rather smaller training set, of just over one million instances. The results for this task are shown in Table 4.16 (highlighted in bold), compared to the reference results and to those obtained from a smaller training set, to minimise differences in results due to size discrepancies.

Again, resampling does not prove beneficial in the context of this task, and accuracy is lower than for either of the imbalanced sets. More strikingly, if the accuracy for the individual determiners is considered, the picture is rather different; the scores are reported in Table 4.17.

Number of training instances	Training data accuracy	Test data accuracy
1,206,787	92.63%	91.32%
reference - 4,043,925	93.26%	92.15%
UNDERSAMPLED - 1,165,428	**87.77%**	**88.58%**

Table 4.16: Determiner results - balanced training set

	Precision	Recall	F-score
a	48.36%	82.27%	60.92%
the	90.11%	74.22%	81.40%
null	99.38%	96.89%	98.12%
macroaverage	79.28%	84.46%	80.14%

Table 4.17: Individual determiner results - balanced training set

The ***null*** case remains apparently easily distinguishable, but now ***a*** is displaying a much higher recall, and ***the*** a much lower one. This may at first appear rather baffling, and counter the initial hypothesis that the indefinite determiner is harder to learn. However, examination of precision scores shows that the great improvement in recall for ***a*** comes at the expense of precision, which is greatly diminished compared to its equivalent in the non-equalised data set; indeed the F-score is virtually unchanged between the two. It seems that having a balanced dataset increases the proportion of times the classifier assigns the label ***a***, which, given its lower frequency in the data, necessarily entails that a number of such assignments will be incorrect – hence the low precision score. The conclusion reached in the similar discussion for prepositions appears valid here, too: it is better to have an unbalanced training set, if this mirrors the distribution of data in the language. This is especially evident in this task, given the importance of privileging precision over recall.

We can also calculate a confusion matrix for the original determiner task, which can be informative in assesssing, among other things, the influence of the ***null*** class over the other two classes. The very high precision score reported for that class would suggest that despite its overwhelming frequency, it is not being selected as the default option in every case for which it is hard to assign a label. The full picture is presented in Table 4.18.

What is immediately evident is that for DAPPER's mistakes involving ***a*** or ***the***, the erroneous choice is almost always the other determiner rather than the ***null*** case. This suggests first of all that the frequency effect is not so strong as to override any linguistic information the model has acquired, otherwise the predominant choice would always be the ***null*** case. On the contrary, these results show that the model

Target det	Confused with		
	a	**the**	**null**
a	xx	92.92%	7.08%
the	80.66%	xx	19.34%
null	14.51%	85.49%	xx

Table 4.18: Confusion matrix for L1 data - determiners

is indeed capable of distinguishing between contexts which require a determiner and those which do not – and indeed that such contexts are distinctive – but requires further fine tuning to perform better in knowing which of the two determiner options to choose[10]. Perhaps the introduction of a discourse dimension might assist in this respect. After all, despite some general syntactic patterns which are typical of one or the other determiner, as discussed in Chapter 3, often the choice between them depends entirely on the discourse context. On the other hand, both Han et al. (2006) and Lee (2004b) attempt to incorporate this information into their feature set. This is not found to bring significant improvements, although it may be that a more sophisticated implementation of it is required, such as relying on co-reference resolution or on semantic representations of the text (as generated for example by Boxer, described in Curran et al. (2007)).

This also makes it harder to offer a detailed error analysis: since in the current implementation of the system the sentences are processed individually and out of order, it is hard to relate them to their original context and therefore understand the extent to which discourse factors might have influenced the choice of the determiner in the original text. This is especially true for the confusion between definite and indefinite determiners. For the vast majority of cases where an incorrect decision has been made by the classifier, the incorrect choice is equally grammatical and acceptable, suggesting that the choice of determiner lies entirely in the discourse, and no other meaningful error patterns can be discerned.

Nor is it much easier to identify clear 'error triggers' in the confusion between the ***null*** case and ***the***. One element that does stand out is the relatively frequent presence (almost 15%) of named entities in the instances of null incorrectly labelled as ***the***. Many of these are assigned the wrong NE tag (e.g. *Dublin, Calais* tagged as organisations) or are not NEs at all (e.g. *trust* tagged as date), which unsurprisingly misleads the classifier into labelling them as requiring the definite determiner. This type of error suggests that the patterns learnt by the model are reliable, but that

[10]Indeed, if only the issue of whether the context requires a determiner or not is considered, precision and recall approach 98%.

other tools used in the process of creating instances for training and testing are perhaps less so. Interestingly, the converse – a high proportion of NEs in the incorrectly tagged instances of ***the*** – is not observed.

This section analysed the performance of DAPPER on individual determiners, together with a confusion matrix showing the types of errors made. It is concluded that while the model can predict with a high degree of accuracy when a determiner is required, the lack of discourse features may be preventing a better performance on distinguishing between the need for a definite and an indefinite determiner. This, together with the fact that the instances in the test set are decontextualised, also hinders a more detailed error analysis. Overall, the initial claim (cf. Chapter 1) that determiners, unlike prepositions, are discourse- rather than lexical item-dependent is supported by these results.

This chapter offered a detailed analysis of the results obtained by DAPPER, taking into consideration various dimensions of variation, such as the technical parameters of the machine learning algorithm and variation within the datasets. The discussion allowed for several insights into the best approach to use in devising such a model, as well as important conclusions on the learnability of individual prepositions and determiners.

Chapter 5

The L1 model: individual feature analysis

This chapter presents a detailed analysis of the contributions made by each feature in the use of the models. The aim is to establish whether the linguistic intuitions followed in designing the feature sets are supported by the quantitative findings of such an analysis, or whether the sets require further refining. The results discussed here seem to indicate that the most important features in both tasks are those relating to POS information and, for the preposition task especially, to the lexical items involved. More rigorous analysis, for example using information gain, is needed to confirm these findings.

5.1 Quantifying the role of features

In Chapter 3, all the feature categories included in the representation of prepositions' and determiners' contexts were described, but there was no discussion of whether they all make an important contribution to the classifier's performance. In this chapter, the extent of this contribution will be assessed. There are two ways of doing this: training a model where one feature category has been removed from the training data, i.e. leaving all others present, and training one where all categories except one have been removed from the training data. In the first case, the variation in accuracy will give some indication of the role played by that feature category. If the figure does not change greatly, it would suggest that the category is likely to play a minor role only. If, on the contrary, accuracy drops, it would indicate that a contribution is made by the category and its presence is important for good performance. Finally, it is also possible that accuracy increases: this would suggest that perhaps this feature category

Feature removed	Distinct contextual predicates	Test data
None	321,377	70.06%
WordNet	321,295	70.10%
Subcategorisation	321,340	70.04%
Multiple modification	321,374	70.01%
Named entity	321,365	69.94%
GRs	321,353	68.61%
POS window	321,104	67.13%
Lexical item	541	57.20%

Table 5.1: Removing one feature category: accuracy

Feature used	Distinct contextual predicates	Test data
All	321,377	70.06%
Lexical item	320,836	60.68%
Modified lexical item	111,494	41.12%
Object lexical item	209,342	37.75%
POS window	273	36.76%

Table 5.2: Using only one feature category: accuracy

is introducing noise rather than assisting the classifier in building a reliable contextual model, and that the classifier might benefit from it being removed altogether.

The converse of this type of experiment, using only one feature category, gives a stronger indication of the role played by that category. Although it is not claimed that preposition or determiner choice depends entirely or mostly on a single factor – indeed the belief is that there is a complex interplay of features at work which is not easy to quantify clearly – it is possible that some feature types are more dominant than others. By omitting all the feature categories except one, a good indication of how much impact a given individual category has when taken independently from the others can be obtained, although of course this does not take into account any interactions among the features.

Together, the results of these two sets of experiments can help refine the model, reducing the impact of noisy and unhelpful features, and ideally improving Dapper's performance, as well as offering new insights into the linguistic dynamics of preposition and determiner choice. In the rest of this chapter, the relevant results for each feature category will be presented.

5.2 Prepositions

The main findings are summarised in two tables, Table 5.1 and Table 5.2, to give an overview of how the results for each type of task compare to each other and to those for the full-featured set, before discussing each feature category in some detail. Not all feature types have been considered for the 'one feature only' type of task. This is because both linguistic intuition and the results of their associated task in the 'one feature removed' task suggest that their contribution, though valid, is highly unlikely to be very large, so that any inclusion in that task would be uninformative. The discussion begins with those features which do not appear in the 'one feature only' table. The table also includes the number of distinct contextual predicates for the various training sets, to show how much the removal of one type of feature affects the size and make-up of the feature space.

5.2.1 Grammatical relations

If the feature category pertaining to the GR information (for relations with both object and modified item) is removed, the table shows that the number of distinct contextual predicates changes only slightly. This small variation only serves to highlight what was already known, i.e. that there is only a small number of GRs relating to preposition occurrence. Such slight changes would lead one to predict that accuracy scores for a dataset from which the GR feature category has been removed should not display much variation from those of the full dataset.

In fact, although accuracy does decrease, it is only by about 1.4%. This suggests that the GR feature, despite being relatively undifferentiated (as there are only a small number of values it can take, and of these an even smaller number are found in the majority of instances), is playing at least a small role in preposition choice. What does this mean in practical terms? From a feature selection perspective, it might be tempting to conclude that one can do without this particular feature category, since its contribution is not so large; furthermore, extracting the information pertaining to it adds a level of processing to the procedure. Howevever, if this stance were adopted for every feature which appeared to contribute even just 1 or 2% to the final outcome, this would fairly quickly sum up to 10 or 20 percentage points being excluded, which is clearly not desirable. Unless this analysis shows a particular feature category to be actively hampering the classifer's performance, or making no change to it whatsoever, it should be included in the feature set, as it may be hard to fully understand the complex relations occurring among the various features.

From a linguistic perspective, the conclusions that can be drawn are less clear-cut. It was noted in Chapter 3 that not all relations occur with all prepositions. If some GR types were predictive of particular prepositions, the absence of this information would make it harder to identify these prepositions: the small drop in performance observed, then, could be due to this. Furthermore, it is worth considering just how much unique information this feature carries, and whether it is possible that some of this information is also inherent in other feature types. This possible redundancy could explain why there is such a small difference between this dataset and the full featured one. An example of this redundancy is the case where if the preposition's object is a noun, it will almost always be in a 'dobj' relation with this noun: this is a duplication of information which expands the feature space but adds little that is new.

On the other hand, other relations do bring something which might otherwise go unrecorded. For example, verbs modified by a preposition can be either in an 'iobj' (indirect object) or 'ncmod' (non-clausal modifier) relation with them. Not only do different verbs appear in different relations depending on context (e.g. verbs of motion vs. stative verbs), but the same verb can appear with either, depending on the type of complement. For example, for the verb *drive* the 'iobj' relation is more frequent when the preposition modifying it is ***to*** (*drive to London*), while 'ncmod' is more frequent when it is modified by ***for*** (*drive for miles*). In this case, the GR information is important as it gives a clue as to what kind of VP is present, which in turn determines what preposition is more appropriate. Indeed, the data shows that the relation 'ncmod', more typical of NPs and denominal verbs, is more frequent in prepositions suggesting staticness such as ***at***, ***by***, ***in*** than in ones typical of movement such as ***from*** or ***to*** (in the latter group for example it occurs in only 24% of instances vs. 80% or more for the others). Conversely, the 'iobj' relation, typical of verbs of movement, is much more frequent in the occurrences of ***from*** and ***on*** (cf. *jump on*) than more 'noun-typical' prepositions such as ***by*** or ***of***. This kind of information is unlikely to be captured by the other features.

It is concluded that although the contribution made by the GR features is not large in terms of percentage points, it is nonetheless valid. The inclusion of this feature type is further justified by the observation that it allows the recording of linguistic peculiarities which might otherwise go unnoticed, and which can be helpful in drawing generalisations and guidelines to preposition use.

5.2.2 Verb subcategorisation frames

The verb subcategorisation frames are another feature category which draws only on a pre-defined, small number of possible values; additionally, as noted in Chapter 3, the resource used for this information has relatively limited coverage, so it is likely that its impact on the classifier's learning procedure is rather small. Indeed, Table 5.1 shows that the number of distinct contextual predicates remains almost unchanged by its removal. As in the previous section, then, little or no change in accuracy is expected.

This is confirmed by the figures in the table, where there is a barely noticeable 0.02% drop in accuracy. Clearly, despite the feature occurring in about a quarter of the data, its coverage is too limited, as noted above, to make a big contribution. The slight decrease in the test data accuracy could be due to the fact that, in the absence of other features not found in the test instances, the model relies on the subcategorisation frames in its attempt to assign the correct label to the instance.

The usefulness of including this feature is debatable. Apart from the scarce impact on performance, it can be surmised that, as the subcategorisation information depends on the verbs being modified, this information is somehow captured by the presence of the lexical item feature anyway, and so is redundant (which could also explain its low impact). It could be argued that with a resource with better coverage, the outcome could be different. However, it is also believed that feature-based descriptions of context can be used to identify subcategorisation frames, and therefore can serve as the basis for creating a new information resource encoding such frames. Having subcategorisation frames serve as part of the context description if there is no independent source for their information besides the context itself might lead to a certain circularity of argument.

In conclusion, while subcategorisation frames are a rich source of information for many kinds of NLP tasks, given the data source used, as noted above, it would appear that they are not best suited to this particular type of task and can perhaps be excluded from the feature vectors.

5.2.3 WordNet

In introducing the features relying on WordNet, it was anticipated that the implementation chosen for this source of information might not be the most informative; therefore, it is particularly interesting to examine the outcome of this task to see

whether this initial prediction is verified or not. In practical terms, their removal reduces the number of distinct contextual predicates, albeit only by a small amount.

The accuracy score on this task displays a small increase only, of about 0.04% more than the full-featured set one. The fact that we see this result, rather than a lower one, supports the hypothesis about the usefulness of the current use of the WordNet information. Including all the available classes may not be the best way to proceed, as it may lead the classifier to find similarities among instances where they do not in fact occur. It is important to establish whether the issue of better performance without the WordNet feature is due to its implementation only, as this would go some way towards answering the question of how useful it is to include semantic features as well as syntactic ones to the contextual representation. Issues regarding homonymy also need to be addressed, to ensure that only the classes relevant to that particular sense or item are included in the representation[1].

If this type of semantic component is necessary, it might be better to experiment with reducing the number of classes given for each lexical item, for example only giving the first one; this is especially relevant for verbs. A first, basic attempt at this involves picking only the first class given for each item, without performing word sense disambiguation or checking whether it is indeed the most prototypical class for that item. With this setup, the result is marginally higher than both the reference result and that for the 'no WordNet' set: 70.24%. This suggests that there may be value to the semantic component, and its contribution could be higher if its use were to be further refined; the current use of the WordNet feature appears counterproductive. It is not clear, however, if the WordNet lexicographer classes are the resource best suited to the model's needs, or if a technique such as clustering of lemmas might be more appropriate: more work is needed to incorporate semantic features in an informative way.

5.2.4 Named entities

The discussion now turns to another feature category which draws on a fixed set of values only: named entities. In describing this element in Chapter 3, it was suggested that there could be some informative relations between particular prepositions and NEs, which this task can test. The number of contextual predicates of course decreases

[1] An example of this is the noun *lap* (example taken from Saeed (2003)), which can refer to a body part or to the distance run by someone, and is associated with four WordNet categories: action, artifact, body part, and stable state of affairs. It is assumed that after disambiguation some of these categories would be discarded.

only slightly as this feature can only take six values for the modified noun and six for the object one.

If the NER is reliable[2], and the hypothesis presented in Section 3.1.2.2 is correct, a certain decrease in accuracy would be expected for this task. In fact, as seen in the results table, there is indeed a decrease, but only of just over 0.1%, somewhat less than one might have been led to expect. There are many possible reasons for this. The most immediate one is the fact, mentioned above, that inaccuracies in the NER may cause certain similarities to be overlooked because of being mistagged.

It is also possible that the linguistic importance of NE for preposition identification has been overestimated. However, this would run counter to our intuitions: as we know, prepositions are used in spatial, locative, and temporal expressions, all of which often include a proper noun. Furthermore, a quick analysis of the data reveals that the distinctive patterns of association between particular prepositions and NEs is confirmed[3]. For example, organisation NEs are most likely to be modified by the prepositions ***for*** and ***of***; the latter is also the one to most frequently modify dates (cf. *in January* ***of*** *2007*).

Overwhelmingly, though, NEs occur as objects rather than heads of PPs. Among these, several strong associations are found. For example, person named entities occur as the object of ***by*** far more frequently than for any other preposition, which is not surprising as, among the NEs, these are the most likely to be involved in agentive roles (e.g. in passive constructions). For similar reasons, ***by*** is also strongly associated to organisation NEs, along with ***at*** (*a meeting* ***at*** *the General Headquarters*) and ***of*** (*the CEO* ***of*** *Google*). Date NEs would be expected to pattern with ***on*** and ***in*** (***on*** *Monday,* ***in*** *June*). While the former is certainly the case, the co-occurrence frequency with ***in***, although higher than with others, is not as high, and in fact is close to that found for the preposition ***for***, which is somewhat unexpected.

Also somewhat surprising are the co-occurrence patterns displayed by location NEs: strong ones are found for ***at***, ***in***, and ***from***, in line with expectations, but not with ***on*** or ***to***, which one would also think of as associated with locations. While the former can be explained by noting that in fact locative complements of ***on*** tend to be common rather than proper nouns (***on*** *the table,* ***on*** *the noticeboard*, etc.), the latter

[2]This is not a trivial point: at the moment no other NER has been tried. The accuracy of this one for English data is reported as just under 85% (F-score) Curran and Clark (2003b), so, as for the WordNet feature, this might be a case of a useful feature being represented in a non-useful way.

[3]For this analysis, the relative frequency of each type of NE among each preposition was calculated. A particularly strong collocation pattern is deemed to be one where the NE occurs in a higher proportion for that preposition than it does in the others.

is more puzzling; perhaps it is a consequence of ***to*** being a very frequent preposition overall, so that associations with NEs account for only a small part of its occurrence patterns. Finally, it is observed that money NEs are extremely infrequent, making it hard to draw any significant conclusions about their patterning, as are time ones, apart from a slightly higher occurrence among instances of ***at***, which is expected (cf. ***at*** *2 o'clock,* ***at*** *midday*).

What stands out in an analysis of this feature's impact is the fact that it is relatively infrequent in the data. In almost nine million instances, NEs occur just over 1.6 million times, less than 20% of instances. Despite their clear correlation with particular prepositions, which could in principle be of some assistance, and is certainly of great linguistic interest, it must be concluded that perhaps their overall frequency is too low for strong patterns based on them to be reliably built by the model, and that the contribution they might make has been overestimated. However, since their inclusion in the model is not an obstacle to accuracy, as proved by the figures, the point made in the discussion of the GRs is reiterated: this feature is kept, on the basis of the claim that several small amounts add up to a bigger contribution to the total score.

5.2.5 POS window

With the POS information, we begin the discussion of those features which have also been used on their own. In the previous chapter, it was argued that while POS features were likely to contribute some useful information which might be hard to capture by other means, they might not be sufficient on their own for a thorough representation of the prepositions' context. Because of this claim, and because of their high frequency for all instances, it is sensible to test their role both by removing them from the feature set and by using them as the only feature category, as this will give as complete a picture as possible of their value. The outcome of these tasks is of interest not just in view of offering linguistic motivation for the need for a more complex feature set than one using just POS sequences, but also in practical processing terms. Since acquiring POS sequences only is clearly simpler and less resource intensive than performing a full syntactic and semantic analysis of the text, should it be found that this category has a major impact on performance, this could lead to a significant revision of the feature choice and to a reduction in the number of more complex ones.

Table 5.1 shows that there are fewer distinct contextual predicates, although still in the same order of magnitude as the full set. Should good results be obtained even

with the removal of this contextual predicate, it would be evidence that strong basic patterns exist and rely on relations with the heads and objects rather than the context of the surrounding POS tags. On the other hand, it might be that the POS sequences contain information pertaining to the one or two distinctive contextual characteristics which differentiate among the prepositions, such that their removal will cause a drop in performance.

The accuracy figures reported in Table 5.1 present a mixed picture. Accuracy is only lower by about 3%, which is not a large amount. However, this also represents the second biggest drop after the lexical item feature category, so it can be argued that of the non-lexical item categories, POS information is the most informative feature. While it is evident that there is information contributed by this feature without which the classifier does not perform as successfully, it is also the case that there are several other informative features present which cause the overall accuracy to not decrease too much.

However, a clearer picture of the extent of the contribution made by this feature can only be obtained by analysing the results of the converse of the task described above, i.e. removing all other features except this one. The figures relevant to this task are presented in Table 5.2. Of course, in this type of task the feature space is vastly reduced: here, for instance, there are only 273 distinct contextual predicates. This adds a further level of interest to the task, as it will be informative to see whether such a reduced number of features can still achieve a satisfactory performance.

Accuracy is around 37%. While still above the baseline by a few points, it is clearly not at all satisfactory. This is not so surprising; after all, the removal of this contextual predicate did not cause a dramatic drop in performance, thus strongly suggesting that it is not of central importance to the classifier, so it is unlikely that this feature on its own can be successful at labeling a large number of instances. This reinforces the belief that this type of feature is best used in conjunction with other sources of information. It appears that POS context, though carrying some weight in determining the preposition belonging to that particular context, is not sufficiently distinctive for this task.

5.2.6 Lexical items

The feature category with the greatest variability is the one relating to the lexical item and POS of the head and object of the preposition, as of course any number of distinct lexical items can occur in these slots. The results of the experiments involving this category may be crucial in establishing the extent to which preposition choice

is predictable. If one assumes that it is idiosyncratic and dependent on individual lexical items, then the removal of this category is likely to cause a significant drop in accuracy. If, however, the belief that there are more general, non-lexical-item-specific patterns involved in preposition selection is correct, a feature set not containing this feature type should still perform reasonably well.

As noted in Section 4.2.1, lexical items account for a large number of low-frequency features. This is also the only feature category to include lexical items, which creates a highly diversified feature set. Therefore, it is not surprising to find that the removal of this category leads to a set of just 541 distinct contextual predicates, which may be too small to successfully classify instances.

As shown in Table 5.1, the removal of this feature causes a sharp drop in performance of over 13%. This figure indicates that although the other features certainly play a big role in preposition choice – otherwise the results here would be even lower – the individual lexical items are certainly a major factor. There are several implications that follow from this. From a machine learning perspective, it means it is unlikely that the information carried by this type of feature can be ignored, which in turn entails remaining tied to a model with a large feature space. From a linguistic perspective, it may have to be concluded that abstract patterns are not sufficient to correctly assign labels to instances, and that, although not going so far as to claim that preposition choice is wholly idiosyncratic, there certainly are some strong lexical item-specific rules which play a key role in the process. This conclusion offers strong support for those theories of language which view language as being composed of fixed chunks and phrases, rather than abstract syntactic frames that we fill each time with the appropriate items (see for example work by John Sinclair, such as Sinclair (1991b); also Hunston and Francis (2000), Cowie (1998), Granger and Meunier (2008)). This is also important from a pedagogical point of view, as it means that students of English may have more success in acquiring the language through such item-based phrases rather than by trying to memorise general, unspecific rules.

This feature category stands out for being the one which causes the biggest drop in performance when it is removed, suggesting it plays a major role in preposition choice. Is this strong effect still valid when it is the only feature category present? The results of this task can give us a clear picture of the extent of the contribution of lexical item features. In preprocessing the dataset for this experiment, it was found that a number of instances had to be excluded because they did not have any lexical item information in their feature set, so that the removal of all other contextual

predicates left them featureless and unusable. A total of just over 100,000 training instances were excluded because of this, and over 6000 test ones.

The number of distinct contextual predicates is high, which points to a lack of uniformity. This, however, is easily explained by recalling the high proportion of lexical item features which occur one time only: if those were discounted, we might have a rather more representative picture of which lexical items have the strongest collocations with given prepositions.

The results of this task offer strong support for the claim that lexical item collocations are the driving force in preposition choice since, as shown in Table 5.2, accuracy figures are only 10% lower than those for the full featured set. It was noted above that lexical items account for the vast majority of low-frequency features, including most of those with frequency 1. It is possible that the presence of so many such features is hindering the classifier's learning, so results for this task when a cutoff of 11 is imposed were also examined. This leads to a reduction in the number of distinct contextual predicates to only 34,455, about 90% less than previously. The effect of this compacting of the data is rather less striking, however, as accuracy only improves by less than 2% (to 62.22%). Therefore, while features which occur only once are indeed the source of some noise, it is likely, as noted in Section 4.2, that the classifier already takes into account the low frequency of these features and therefore assigns them a low weight, so a noticeable difference in performance is not found.

Overall, these results give us two key pieces of information: firstly, that lexical items are a crucial component for a successful system, and secondly, that on their own, however, they are not sufficient to achieve the highest possible accuracy. The first point is proven by seeing that performance when this feature is removed suffers, and the other features cannot fully compensate for its absence; and that on their own, lexical items account for a large part of the classifier's success. As regards the second point, it is noted that, as just mentioned, a set containing the other features, but not the lexical items, can also achieve scores which are higher than the baseline. Furthermore, as the full feature set remains higher-performing, it is clear that there is a valuable contribution that other aspects of the context make, and that preposition choice is not entirely lexically driven, since a combination of different feature types may be the optimal solution for this task. Finally, it is noted that the important role that lexical items are found to have justifies the choice of carrying out syntactic analysis of the text: it must be ensured that the lexical items identified as being part of the PP are those that do indeed belong to it. As previously discussed (Chapter 3), it is unlikely that this would occur if relying on linear sequences of words only.

It is worth probing further into the issue of lexical items to establish whether head and object items are equally important for the model, or whether one has more weight than the other. Heads would be expected to be more determining, since, as observed (Section 3.1.2.1), most things can appear as objects of a preposition, but stricter restrictions seem to be in place as regards the heads. Indeed, this is confirmed by observing the different number of distinct contextual predicates in each set (cf. Table 5.2): there are almost 100,000 more such pairs in the 'object only' set than in the 'head only' one. As for the previous task, it is found that not all the training instances can be used because several remain featureless; in both cases however the figure is over 8 million.

As far as the actual results are concerned, accuracy for these sets is lower than that of the set with all lexical items included, but not as low as the 'POS only' set. Analogously to the previous task, whether singleton features were introducing too much noise was also checked, by running the task with a cutoff point of 11. This brings, again, a sharp reduction of the feature space, but only a small improvement in accuracy, most noticeably for the 'head only' set (head only: 45.08%; object only: 39.41%), which could be further evidence of the greater importance of heads over objects.

From the results of this task, it must be concluded that while heads and objects play different roles within preposition choice, and objects may be less predictable than heads, both are clearly necessary for successful class assignment. From a pedagogical perspective, this implies that in learning about preposition use, it is best to focus not just on head+preposition combinations, but also on preposition+object ones.

5.2.7 Multiple modification

Related to the previous feature is the 'more than one' category, which, as discussed in Section 3.1.2.1, records whether the nouns and verbs modified are modified by more than one PP, and also whether the verb modified has a direct object in the context under consideration. There it was noted that it was not clear how useful this information would be, and that it is possible that certain prepositions are more likely than others to occur in complex NPs with multiple PPs, or that certain nouns or verbs are more likely to require a particular preposition when heading such complex NPs. Therefore, it is of particular interest to examine the effect the removal of this feature will have on accuracy.

The results, however, are not conclusive: there is in fact a slight decrease in accuracy, suggesting that the information contributed by this feature, though to a

Features used	Distinct contextual predicates	Test data
All, no cutoff	321,377	70.06%
All, cutoff 11	34,876	69.79%
POS and Lexical item, cutoff11	34,718	65.20%
All except WordNet, subcat	321,258	70.03%

Table 5.3: Preposition combination models: accuracy

small degree useful, is not in fact of central importance. As this category is related to the lexical item, it seems appropriate to also try a variation on this task whereby both the lexical items and the 'multiple modification' features are removed. This gives slightly lower results: 56.79%. In other words, although the effect of this information is not a major one, it certainly has an impact, especially when removed together with the lexical item feature, and so it is worth retaining as part of the feature set.

5.2.8 Combination models

The discussion above had the aim of assessing the contribution made by each feature category to the model, both for the purpose of gaining linguistic insights into the process of preposition choice, and, from a more practical point of view, to establish what the best possible feature set for the task could be. Having noted that none of the features selected are impeding a better performance – since the absence of none of them causes a dramatic improvement in accuracy – the question arises as to whether this is the best that can be achieved at this task. From the results of these investigations, it may be possible to come up with a combination of contextual predicates which could potentially yield the highest accuracy and minimise noise. The relevant results are summarised in Table 5.3.

The experiments discussed in the previous sections suggest that the two biggest factors in preposition choice are POS and lexical item information, and that accuracy with these features improves when a cutoff of 11 is imposed. Therefore, the first combination model uses these parameters; Table 5.3 includes for comparison the standard reference results as well as those with the full set and cutoff 11. Taken together, these two categories give a higher accuracy than when taken individually, although one that is not much higher than the lexical item only set. Furthermore, the fact that this figure is only a few points lower than that of the full-featured set at cutoff 11 confirms the claim that these two are the main factors in governing preposition usage, although not the only ones, as evidenced by the difference in accuracy scores.

However, as all features apart from WordNet caused a drop in performance when removed, it was concluded that most of the 'minor' contextual predicates were worth

Feature removed	Distinct cont. predicates	Test data
None	185,959	92.15%
POS	185,686	76.85%
Head noun	29,412	90.74%
Adjective info	157,417	91.30%
PP info	185,402	91.78%
Named entity	185,953	92.14%
Relative clauses	185,958	92.14%
WordNet	185,933	92.18%
Other	185,958	92.15%

Table 5.4: Removing one feature category for determiners: accuracy

retaining even if each only made a contribution of one or two percentage points. Subcategorisation information was also found to make a barely noticeable contribution to the total. It would therefore seem that the best course of action is to use a model which excludes only the abovementioned two feature categories. As can be seen from the last row of the table, the results of this model are, contrary to expectation, not better on test data, and even lower when a cutoff of 11 is imposed (69.85%). They are, however, slightly higher than the full-feature set with cutoff 11, suggesting that a good choice of features might include all features, even *hapax legomena*, but no WordNet or subcategorisation ones, at least in the implementation used at the moment.

This section has offered a detailed analysis of the role played by each feature in preposition choice, as quantified by the effect of their presence or absence on the accuracy results. Several insights of a linguistic, pedagogical, and practical nature were given. It is concluded that despite the need for various types of syntactic features, the main driving force behind preposition choice are the lexical items present in the PP, a consideration which ought to inform work undertaken in the fields of both NLP and L2 instruction.

5.3 Determiners

Analogously to the preposition data in the previous sections, here a discussion of the impact of the various features on determiners is presented. Table 5.4 summarises the results for the removal of each category. The final row, 'other', refers to the following contextual predicates whose removal does not affect the results: presence of quantifier, of possessive modifier, of cardinal number, and of 'existential there' phrases.

5.3.1 Minor features

The discussion begins by looking at the just-mentioned group of contextual predicates which are not found to have an impact on determiner choice. In Chapter 3 it was noted, with regard to them, that while there are some grounds for believing they could be an indicator of determiner choice, these are inconclusive and the outcome of this task would be necessary to confirm these intuitions. It may be that these contextual predicates are not very frequent overall. Indeed, in a feature space of over 52 million features, these four features represent only a small fraction of the total, which would lead the classifier to assign them only minor weights in building its model. The frequencies are as follows:

- Possessive modifier - 218,047
- Cardinal modifier - 186,608
- Existential there - 20,915
- Quantifier modifier - 10,048

It is therefore not surprising that their absence hardly has an effect on the classifier, as furthermore their low frequency is accompanied by the fact that the presence of these features in the determiner's context is noted by a single contextual predicate such as '`ModbyPossessive`', '`ModbyQuantifier`', and so on, without further distinguishing lexical information. The results of these experiments show that the importance of the information carried by these features for determiner choice may have been overestimated. Since their frequency in the data is so low, it cannot be conclusively stated if this is due to misjudged linguistic intuitions or is simply the effect of low frequency. Perhaps in principle, these features can be used as guidelines for human learners to identify a need for determiner presence or absence, but it is unlikely that a classifier can rely on them in class assignment.

However, the frequency of these features is not so low that one cannot refer to it to check whether the linguistic intuitions put forth are valid or not. The relative frequencies of the four features within the data are examined to see if there are particularly strong correlations between a given contextual predicate and determiner. Some of the predictions are solidly confirmed. For example, over 99% of cases of the 'modified by possessive' feature occur in instances of ***null*** determiners, as expected. The 'existential there' feature shows a strong relation to ***a***, with almost 50% of its instances occurring with that label, which is also in line with our intuitions. Equally,

it was hypothesised that predeterminers would occur mostly in phrases where a determiner was also present, and this is indeed the case: the 'modified by quantifier' feature occurs overwhelmingly in instances of ***the*** (over 70%) and frequently also in instances of ***a*** (20%).

Finally, the 'modified by cardinal' feature exhibits a partly puzzling behaviour; while over 75% of its occurrences are within the ***null*** class (cf. *three boys, seven lizards*, and so on), there is also a higher than expected number of them in the ***a*** class, which should not occur (cf. **a three boys*). A quick overview of the data reveals that the majority of these are phrases where the noun is a word such as *year, hour*, or *day*, suggesting that these are misanalyses of phrases like *a two hour journey*, *a three year ban*, and so on, where the determiner has been associated with the incorrect noun – the error therefore lies not with the classifier but with the parser. This brief analysis shows that while not all the features chosen may be of use to the classifier, they can yet yield informative linguistic insights and offer quantifiable support for the hypotheses put forth in Chapter 3.

As is clear from Table 5.4, two further features ought to be included in this section by virtue of their scarce effect on accuracy, namely **relative clauses** and **NEs**. Like the features above, it was hypothesised that the presence of **relative clause modification** would be a strong clue to the need for a determiner. The accuracy figures seem to belie this, as the removal of this contextual predicate only brings a small decrease. However, it, too, is relatively infrequent in the data (107,896 occurrences), so it is not possible to draw firm conclusions about its usefulness for the classifier. From a linguistic point of view, it was suggested relative clause modification is especially frequent in NPs with a definite determiner. The data, however, undermines this claim: this feature is found often in all three cases in proportions varying from 20% to 46%, with a slight preference for ***a***. It is therefore unclear whether, were it more frequently found in the data, this feature would prove to be informative, as no strong association with any one determiner is found.

NEs are rather more frequent, occurring about 774,000 times, so their scarce impact on accuracy is more surprising. It is not clear if the lack of effect on the classifier is due to this feature not being informative, perhaps because of a failure of the NER module, as discussed above, or to inherent linguistic reasons. In the main, it is found that the predictions about occurrence of determiners with NEs are confirmed. For example, the great majority of person (96%) and location (84.5%) NEs occur with the ***null*** case. Organisation NEs were also expected to display more variability between definite and null case, and indeed this is observed in the data:

these NEs occur in instances of ***the*** 35% of the time, and in those of ***null*** 63% of the time. Times and dates, too, were thought to be less strongly associated to just one class of determiner, and this is reflected in the data, with ***null*** cases having 75-77% of these NEs and ***the*** 18/20%. Finally, as regards money NEs, these are extremely infrequent (3200 instances). They, too, show a clear preference for the **null** case, but there is also a significant minority occurring with ***a***, which refers to expressions such as *a dollar, a penny, a pound*, and so on. Although it is surprising to see that NEs have less impact than expected on accuracy, it is encouraging to find concrete proof of our intuitions about their behaviour with determiners.

5.3.2 Prepositional phrases

In Chapter 3, it was hypothesised that nouns involved with PPs, especially as their heads, are more often found with a definite article, and that therefore the presence of this feature could be informative for the classifier. PP information is a relatively more frequent contextual predicate, with 'object of preposition' occurring almost 2 million times and 'modified by preposition' almost 1 million times. As these contextual predicates also contain some lexical information[4] (stating which preposition is involved), naturally their removal causes a decrease in the number of distinct contextual predicates, of about 500. It may be the case that the presence of a PP is the only way to distinguish between an NP needing a definite determiner and one that does not, so it is not clear that their removal will necessarily be beneficial.

The removal of this contextual predicate causes only a minor drop in accuracy, of less than 1%, so perhaps, despite its frequency, and the perceived importance of its role within the NP, it is not as determining as previously thought. A closer look at its distribution across the three classes reveals that its associations with the various classes are not as strong and unequivocal as expected; this is likely to be the reason for its minor contribution to the development of the model. Indeed, 'modified by preposition' occurs with roughly similar frequency (ca. 40%) both in instances of ***null*** and ***the***, thus belieing the notion that it is a hallmark of a definite determiner, while 'object of preposition' shows some preference for ***null*** (62%) and ***the*** (29%), which is more in line with the observations on it being a less certain predictor of determiner choice.

[4]Of course there is the same danger here as in the prepositions task, namely that a preposition is taken to be anything that the POS tagger has labelled as a such: currently there are no filters in place to exclude mistaggings.

Overall, it is not clear if PP involvement is a useful feature for the classifier, since its absence does not have a detrimental effect on accuracy, and it has been shown not to have strong correlations with any particular determiner class.

5.3.3 Adjectival modification

Adjectival modification is expected to play a role in determiner choice, especially where the adjective is in the superlative case. The removal of this contextual predicate noticeably reduces the number of distinct contextual predicates, by almost 30,000 (we recall that the actual adjective present is given as well as noting the fact that there is adjectival modification). This may be because the majority of instances of this contextual predicate are singletons so, contrary to the initial hypothesis, their removal might not affect accuracy as much. Indeed, there is a drop in accuracy of only less than 1%, so it must be concluded that this category is not as strongly predictive as expected.

This feature is not very frequent, occurring about half a million times. Of these, the vast majority are adjectives in the base case, which occur about half the time with ***the***, suggesting that perhaps adjectives are used slightly more often to pick out a specific entity, and therefore require the definite determiner. Comparatives, on the other hand, are more evenly distributed between definite (ca. 40%) and indefinite (ca. 35%) cases, which is not surprising given that both phrases such as ***a*** *better thesis* and ***the*** *better student* appear frequently in the language. Finally, the predicted behaviour of superlative adjectives is solidly confirmed, as almost 95% of those instances occur in the definite determiner class. This brief analysis shows that the intuitions about the role of adjectives are confirmed by the data; however, the frequency of the relevant features may be too low to be of great use to the classifier, explaining the scarce change in accuracy.

5.3.4 WordNet

Before discussing the two last major features of this set, the role of WordNet features for determiner choice is addressed here. In the prepositions task, WordNet categories were found to be a hindrance to good performance, as their removal brought about a slight improvement in accuracy. It was hypothesised that this was due to the fact that all possible categories for a word are listed, which could lead to unnecessary noise. The problem seemed to be more salient for verbs than nouns, however, as verbs seemed to be overall assigned to more, and sometimes all, of the available categories.

So it does not necessarily follow that the removal of this category will have an equally positive impact on the determiner data.

There is a small variation in performance, in that accuracy improves by 0.02%. While this is hardly dramatic enough to warrant the claim that the WordNet features are a serious problem for the classifier, it is clear yet again that at least in this form they are not making a useful contribution to the model, so their presence in the feature set may have to be reconsidered, or a different way must be found to include the information captured by this feature.

As done for the prepositions, the task with the selection of one WordNet class only is attempted – again by simply choosing the first one present. Here, too, there is an improvement over the reference results, albeit a much smaller one, of only 0.03%. This may be further evidence that selecting just one WordNet class is better than selecting all of them, but, again, a more sophisticated integration of this semantic feature needs to be devised.

5.3.5 POS window

In the prepositions task, this category was found to be one of the more informative feature categories, so it will be interesting to see whether a similar effect is found for determiners. Table 5.4 shows that there is no great variation in the number of distinct contextual predicates. The accuracy for this task shows a noticeable decrease, of about 15%, suggesting that POS tags may play an important role in this task. To better assess this claim, an experiment where POS features are the only feature category used was also run, as previously done for the prepositions. The results for this task are presented in Table 5.5.

Accuracy is relatively high, only 3% lower than the full-featured set. This suggests that the lower accuracy observed in the previous task is indeed attributable to the importance of the POS features. It would appear from these results that POS on its own, without any further syntactic processing, can go a long way towards correctly predicting determiner choice. How can this be accounted for? The most likely explanation is that several of the characteristics of the context recorded by the other features are also captured by the POS tags, so their absence here is not so much of a problem. On the other hand, if this were the only reason, then the results for 'no POS features' should not have been so low, as the relevant information would have been present anyway – albeit in a more differentiated and fragmented form than as captured by POS tag sequences, which could be a crucial difference.

Feature used	Distinct cont. predicates	Testing
All	185,959	92.15%
POS window	273	89.24%
Head noun	156,547	69.16%
Count/mass	3	53.13%

Table 5.5: Using only one feature category for determiners: accuracy

5.3.6 Head Noun and tag

It is generally agreed (cf. Chapter 3) that one of the most, if not the most, important factors in determiner choice is the head noun itself, both its lexical item (because of properties associated with it such as countability) and its number. Therefore, accuracy is expected to decrease dramatically when this feature category is removed – which involves the removal of both the lexical item and its POS tag (which, we recall, indicates both number and whether it is a common or proper noun). The number of distinct contextual predicates is greatly reduced in this task, confirming that there are many singleton or low-frequency instances within this category, as of course there is almost no limit to what can be used as a noun in the English language.

Accuracy obtained is somewhat striking. If the claim that determiner choice is strongly lexical item-dependent were to hold, performance on this task should be rather low. Instead, there is a drop in accuracy of less than 2% only. This result suggests that the role played by more general contextual patterns, as opposed to individual lexical items, may be much stronger than one might think. This is encouraging for the development of NLP tools, which can be more generally applicable if not tied to lexical items, and of assistance in a pedagogical framework, as it means there may indeed be simple guidelines that learners of the language can rely on in most cases.

As for the previous example, the best way to fully assess the role of this feature category is by looking at the results when it is the only category present, as summarised in Table 5.5. Although these figures at first glance do not seem so low, they are in fact only a few points higher than the baseline, and much lower than what can be achieved with other features. While using head noun features only could be of some use when more complex features cannot be extracted, it is clear from the results presented in this section that they are not sufficient on their own to obtain a satisfactory performance. It can be hypothesised that the other features in the set to an extent fulfil the role of giving information about the discourse dimension and hence are important for a successful assignment of class labels.

Features used	Distinct cont. predicates	Testing
All, no cutoff	185,959	92.15%
Head noun and count/mass	156,550	69.27%
Head noun and POS	156,820	90.97%
Head noun, count/mass, POS	156,823	90.98%
As above plus NE, adj, pp	185,928	92.17%

Table 5.6: Combination models for determiners: accuracy

Another possibility is that the absence of the head noun information is compensated for by another feature which has not been discussed so far, namely whether the noun is **count, mass**, or **either**. The quantification of the contribution made by this feature is therefore attempted. Accuracy is unchanged, suggesting this feature may not be as strongly predictive of determiner choice as expected. It is fairly frequent, occurring almost 3 million times altogether, that is in almost every instance (proving that the resource used for the extraction of this feature has very good coverage). However, there are no very strong association patterns between particular determiners and noun types: apart from mass nouns, which occur with the ***null*** case around 74% of the time, the other two types are more evenly distributed. So it is not as surprising as initially thought that the absence of this feature does not cause particular problems for the classifier.

To further support this conclusion, this feature is also tested in the 'one feature only' task, as reported in the table. In this setup, the dataset only has three distinct features, so it is not so unusual for accuracy to be so low, lower even than the baseline. Clearly, then, count/mass information alone is too generic for successful class assignment[5].

5.3.7 Combination models

The discussion above assessed the role played by each feature category in the development of the model, and noted that several contextual predicates seemed to make little or no contribution to the outcome. As for the case of the prepositions, it must be examined whether a better performing model can be obtained by excluding those features, which would also give gains in terms of processing efficiency. Results from experiments with various combinations of features are summarised in Table 5.6.

Again, the results observed in the previous sections are relied on to select the features most likely to be of use for this task. Firstly, the head noun and count/mass

[5]Nor can it be reliably used as a guidelines for learners of the language: the only firm rule they can rely on is that most of the time, a mass noun will not require a determiner.

information only are used, to find conclusive evidence as to their role in the task: we can see that even when combined, accuracy does not change much. This confirms the claim that this information on its own is not sufficient for the task. The addition of the POS context, shown in the previous section to be the most important factor in the process, does indeed bring the accuracy scores to within 1.5% or so of that of the full feature set. This not only reinforces the claim that POS information is of central importance, but also that there is great value in having a variety of features. It is clear that determiner choice *does* depend on an interplay of features, and while some may be more significant than others, they are all necessary to achieve the best results. For the final task, the three features which were shown to have an impact, albeit minimal, on accuracy are added to the set: NE information, PP involvement, and adjectival modification. Indeed, with their addition the results are not only higher than the ones with head noun and POS information only – evidence that little contributions do add up – but also marginally higher than the full-featured set.

The results observed in these tasks show that although several contextual features may be of linguistic and pedagogical relevance in discussing the properties of NPs and determiners, they may not always be as relevant in developing a successful and efficient model and should therefore not be included in a feature set for this task. POS tags seem to play a central role in determiner choice, together with information about the head noun itself. However, syntactic analysis should be considered a desirable addition to the creation of the feature set, as it ensures that the correct relations between elements of the NPs are captured.

This chapter presented a detailed examination of all the features used in developing the preposition and determiner models. A variety of experiments attempted to quantify the individual contribution made by each. The results of this analysis offer a starting point for reflections on linguistic phenomena of use to NLP researchers, linguists, and language teachers.

Chapter 6

Dapper at work: application to L2 text

This chapter describes the application of DAPPER to its intended target, L2 text. Section 6.1 introduces the Cambrige Learner Corpus, the resource used for the evaluation, and highlights some issues which may arise when using DAPPER with L2 text. Section 6.2 reports the procedure used to extract suitable data for testing from the corpus, and Sections 6.3 and 6.4 give the results obtained on these tasks for prepositions and determiners respectively. Finally, Section 6.5 summarises the main issues encountered in using DAPPER on learner language.

6.1 The Cambridge Learner Corpus

This section provides an overview of the composition and structure of the Cambridge Learner Corpus, before going on to address some of its characteristics which may be considered problematic when NLP tools are applied to it.

6.1.1 A description of the corpus

As the source of authentic and representative L2 data for evaluation, a 2 million word subset of the Cambridge Learner Corpus (CLC) is used[1]. The CLC is a corpus of contemporary written learner English, which currently stands at over 30 million words[2]. It is developed jointly by Cambridge ESOL ('English for Speakers of Other Languages') and Cambridge University Press. The corpus contains material from a wide range of students, both as regards their L1 and their proficiency: there are

[1]This has been made available to us by Cambridge University Press, whose assistance is gratefully acknowledged.

[2]See http://www.cambridge.org/elt/corpus/learner_corpus2.htm.

scripts from 95,000 students from 190 different countries, with 130 different L1s. The essays represent answers from several kinds of ESOL tests, from beginner level to the more advanced *Certificate of Proficiency in English*.

This data is annotated in two ways. Firstly, for each script, information regarding the student's L1, nationality, age bracket and proficiency level is recorded. Secondly, each script is error-tagged, as discussed in Section 2.1.3. Within the 2 million word subcorpus used, there is still a good range of variation among the characteristics of the learners present. Over 120 different nationalities are represented, from all continents; these include 60 languages from several groups: Slavic, Romance, Indo-Iranian, Dravidian, South-east Asian, Chinese, and so on. This is an important factor because it is often said that speakers of different L1s tend to display different error patterns with regard to prepositions and determiners. In using data which represents several different L1s as a testbed, there is a higher likelihood of developing an application that can be of use to a wider range of learners.

As for the level of proficiency displayed, all exam types found in the whole of the CLC are also represented here; these include the *Business English Certificate*, the *Certificate of Proficiency in English*, and the *Key English Test*, as well as data from *International English Language Testing System* (IELTS) exams. This means that as well as a range of skills, one can also expect to find a range of topics in the learner texts, and therefore a range of vocabulary and syntactic structure. Common exam questions require the writing of a business letter or report, essays which recount personal experiences or argue in favour of a position, or more informal letters to friends.

For most students more than one text is included, giving 16,844 distinct answers overall, with an average length of 120 words per text (although of course some of the essay-type ones will be longer, and some of the letter-type ones rather shorter).

6.1.2 Possible issues in using NLP with L2 data

Chapter 2 introduced some of the general issues which can cause problems when using conventional NLP tools on non-native English text. It is likely that some of these problems will affect this task; this section briefly considers what kinds of obstacles might be encountered.

At the heart of Dapper's feature vector construction is the processing of the output of the C&C tools pipeline. These tools have been trained on correct English newspaper text, more specifically the Wall Street Journal, which contains articles on economics and business topics. While, as mentioned above, business writing is one

of the skills that can be assessed by the Cambridge examinations, the style of those scripts is unlikely to be very similar to that of the WSJ; furthermore, other text types are present in the CLC, too. It is possible that this shift in text type will have an adverse effect on the parser, and lead it to output parses which do not correspond to the structure of the sentence. On the other hand, an informal analysis of the parsed BNC data showed that this problem was not encountered often on that task. Additionally, the syntactic structure of the learners' writing will likely be rather more simple than that of newspapers and so should not prove a challenge for the parser in any case[3].

Errors in word order could pose a problem for the tagger or parser, too. Since the POS tagger relies on a statistical model whose features are the lexical items and preceding POS tags in a window around the target, certain words in the sentence may be tagged as the POS it expects to find, rather than the correct one. This in turn may lead to incorrect parses, relation assignments, or feature extraction. An example is the sentence *I can you also send a map of London*, where *can* is analysed as a main verb rather than an auxiliary and assigned *you* as its direct object; in turn, *send* is not recognised as a ditransitive verb, and the overall resulting structure of the sentence is severely distorted. Parsing can also be affected by agreement errors, also often found in learner writing. If the tags have been appropriately assigned, there may be a subject and verb which disagree in number, which may not be recognised as belonging together. Despite these problems, a quick overview of the parser output shows that in the main POS tags, syntactic structure and GRs are correct; the contribution made by these types of analyses is rich enough to outweigh any errors that might appear in the data. This decision is supported by the findings reported in Kakkonen (2007), where the CCG parser outperforms three others (Link Grammar Parser, Stanford Parser, StatCCG) in parsing ungrammatical sentences containing between one and three misspelled words.

More significant problems are likely to arise at the lexical item level, both because there are several different elements that are susceptible to problems, and because, as seen in the previous chapter, many lexical items play an important role in the model's performance. One of the most evident issues is that learner writing contains a large number of spelling mistakes. Spelling mistakes which lead to non-existent words negatively affect various components of the model. The lexical item may be

[3]Assuming, of course, that it is essentially well-formed; as shown later, there are cases where a combination of errors makes it impossible for the parser to assign a correct parse, or indeed any parse.

incorrectly stemmed, or not stemmed at all. Any information associated with that (correctly spelled) item will not be retrieved, including the use of the item itself as a contextual predicate. An example sentence with a preposition error illustrates this point, with the relevant word in bold:

(4) John understood straightaway the **reson** [**reason**] of her visit.

The misspelled word in (4), *reason*, is one which is not only rather frequent in English, but also has an especially strong collocational tie with the preposition ***for*** – indeed, it is found as the head of a PP in the L1 data 9800 times. However, the link to this determining information is lost as the misspelled word cannot be matched to its correctly spelled equivalent. In analysing the performance of DAPPER, this is an important factor to take into account.

A related issue to this is the presence of misspelled words which result in another English word, some examples of which are given below, with the correct item included:

(5) I would like all the members to notify **stuff** [**staff**] in their section.

(6) ... The first **flour** [**floor**] of a house, with all modern facilities...

(7) Excepting the **brakes** [**breaks**] between each course, the seminar was well organised.

With these items, the problem is that the information accessed by DAPPER will be that pertaining to the incorrectly spelled but legitimate word. Furthermore, while the previous misspelling issue can be addressed by running a spellcheck on the text before submitting it to the system, this type of misspelling would not be recognised by this method. The error coding is only of partial assistance: these errors are tagged sometimes as 'spelling errors resulting in a legitimate word' ('SX') and sometimes as 'replace noun' ('RN'), analogously to those cases where a completely different lexical item has been used inappropriately (e.g. *coming* for *arrival*, *firm* for *office*, *model* for *way*), so cannot always be reliably identified.

In this section, the CLC was introduced, along with some errors which are expected to impair DAPPER's performance. In the course of this chapter, a detailed analysis of the results obtained is provided. This will establish whether these characteristics of learner writing prove problematic for the system, or if instead other unanticipated problems arise.

6.2 Creating a test set

This section describes the method used to extract the corpus data for the evaluation of DAPPER, and what preprocessing steps need to be carried out to make the data suitable for use.

The aim of DAPPER is to reliably detect preposition and determiner errors in L2 writing, so the best test of its success is the assessment of its performance on sentences containing such errors, noting how well it does in recognising the errors and suggesting an appropriate, more idiomatic alternative. However, this is not sufficient: it must also be ensured that it does not raise false alarms, that is, it does not flag the presence of an error where in fact there is none. This possibility, not unlikely given its imperfect performance on L1 data, is made even more likely by the factors liable to impair NLP tools, as described in the previous section. Avoiding such false alarms is of even greater importance here than in dealing with L1 data, because it would harm a learner's confidence, and perhaps progress, to be notified of non-existent errors. Therefore, in creating this test set both correct and incorrect prepositions and determiners are included. In both cases, correct instances outnumber incorrect ones, as is normally the case in learner text: despite the frequency of these errors types, they still represent a fraction of all occurrences of the two POS.

In identifying occurrences of incorrect items, the error tags present in the corpus are used. For prepositions, at the moment DAPPER is only trained to recognise appropriate usage of a set of nine prepositions, and only where a preposition is required. So, only preposition errors of the 'RT' type, i.e. selection errors, are considered: those where a preposition is needed, but the one chosen by the student is incorrect. Additionally, it must be ensured that for all these instances, both the incorrect preposition and the suggested correction are part of the set of nine of which the system has knowledge, otherwise it would be impossible for it to process them successfully. A typical example of this is *you might be interested **at [in]** giving a talk*. At this stage, text from all levels of proficiency is extracted. In this way, 1116 test instances of incorrect prepositions are obtained; 5753 correct instances are also extracted.

The determiner component, on the other hand, is capable of dealing with all three types of errors because it includes the ***null*** class among its labels. The three error types are: 'MD', missing determiner (where it would recognise that ***a*** or ***the***, rather than ***null***, are required), 'UD', unnecessary determiner (conversely, recognising that ***null*** rather than ***a*** or ***the*** is required), and 'RD', wrong choice of determiner (recognising that an indefinite determiner is more appropriate when an indefinite has

been used, and vice versa). Sentences which contain at least one of these error codes are extracted, yielding 475 test instances of incorrect determiner use.

Collecting instances of correct use is somewhat trickier than for prepositions; while of course it is trivial to search for occurrences of ***a/an*** and ***the*** in unannotated text, it is not so easy to search for a non-visible determiner for bare NPs *before* the text has been POS tagged (which it has not, at this stage), as there will be nothing to search for. However, since bare NPs are the majority in the language, it is assumed that most sentences extracted will contain at least one such NP. Indeed, by proceeding in this way 1760 instances are collected, within which the three determiners have a similar distribution to that observed in L1 data.

To obtain the instances used by the system, the sentences extracted undergo some further preprocessing steps in addition to those used to treat the L1 data (cf. Chapter 4). The sentences are stripped of any residual XML markup relating to the text's overall organisation (e.g. paragraph and script breaks), as well as both the XML tags for the error codes and the corrections inserted by the annotators. This is because DAPPER must be presented with text of the same kind it would receive if a student were inputting sentences directly, free of any markup introduced by the corpus annotators.

6.3 Prepositions

As described in Section 6.2, the prepositions test set contains 6869 instances for DAPPER to mark as correct or incorrect. Taken as a whole, accuracy on the task – the number of times a correct instance is marked as error-free, and an error is recognised as such – is 70.2%. However, this figure is misleading as the proportion of instances which are both recognised as an error *and* given the appropriate correction is not high, as will be shown in Section 6.3.2. To gain a clearer understanding of issues arising on the two types of data, these will be treated separately.

6.3.1 Performance on correct data

The discussion begins with an analysis of DAPPER's performance on the correct L2 preposition instances, as this provides the ideal context for direct comparisons with the L1 task and text types. It is important to stress that although these instances contain no *preposition* errors, they may well present errors involving other POS or structures. The use of this data can help us understand the extent to which the system can be applied to types of texts and languages different from those seen in

Correct instance, found	69.00%
Baseline on correct data - of	17.17%
Baseline on correct data - in	28.71%
Error, found and well corrected	39.51%

Table 6.1: Accuracy on L2 data - prepositions

training. The setup of this task is the same as that for the L1 data: feature vectors are created for instances of correct preposition use, and the system is required to assign the original class label to them.

5753 instances are submitted to the system, of which, as shown in Table 6.1, 69% are accurately labelled. Three models were tried on this task. The generic one, using all features, achieves accuracy of 69%, as noted. The model in which all but one WordNet class is removed, found to bring a small improvement on the L1 data (cf. Section 5.2.3), does not have much effect here: accuracy is 69.09%, while the one which does not use any WordNet or subcategorisation features at all achieves 68.30% accuracy. In light of the negligible difference between the two better performing models, the results from the full-featured model are taken as the reference in all discussions in this chapter. If the same baseline as the L1 task is used, namely always choosing ***of***, the most frequent preposition according to L1 figures, accuracy would be rather low. However, in this dataset the most frequent preposition is ***in*** rather than ***of***, so a fairer baseline might use that as its default choice. In this case, the baseline accuracy is higher, and the improvement of the model over the baseline is comparable to that observed in the L1 task.

The figure of 69% is an encouraging result as it is only 1% lower than that achieved on the BNC data. A loss of accuracy in moving to a different domain can be expected, but the loss here is quite small, pointing to a robust model which is not too tied to a specific domain. Furthermore, it can be taken as evidence that the L2 nature of the texts is not, at least at first glance, creating too many problems for the NLP tools. In the related literature, there is – as far as it could be ascertained – no similar analysis of the accuracy of such applications on error-free text only, so this result is difficult to contextualise; however, it is believed that the issues raised may make a positive contribution to the study of NLP tools applied to L2 data.

A relatively good performance on correct instances is important in view of this application's target audience of learners; as discussed, it is important to minimise the rate of false alarms as far as possible. It is also interesting to look at the results more closely, including precision and recall scores for individual prepositions, as this will

	Proportion of training data	Precision	Recall	F-score
of	27.83% (2,501,327)	74.28%	90.47%	81.58%
to	20.64% (1,855,304)	85.99%	81.73%	83.81%
in	17.68% (1,589,718)	60.15%	67.60%	63.66%
for	8.01% (720,369)	55.47%	43.78%	48.94%
on	6.54% (587,871)	58.52%	45.81%	51.39%
with	6.03% (541,696)	58.13%	46.33%	51.39%
at	4.72% (424,539)	57.44%	52.12%	54.65%
by	4.69% (421,430)	63.83%	56.51%	59.95%
from	3.86% (347,105)	59.20%	32.07%	41.60%
macroaverage		63.67%	57.38%	59.68%

Table 6.2: Individual prepositions results - L1 data

	Proportion of test data	Precision	Recall	F-score
of	17.17% (988)	67.66%	89.57%	77.09%
to	20.35% (1171)	88.86%	78.31%	83.25%
in	28.72% (1652)	69.85%	75.30%	72.47%
for	10.95% (630)	66.67%	57.46%	61.72%
on	6.12% (352)	57.29%	46.88%	51.56%
with	4.64% (267)	58.75%	35.21%	44.0%
at	5.65% (325)	45.28%	57.54%	50.68%
by	3.06% (176)	45.86%	40.91%	43.24%
from	3.34% (192)	57.75%	21.35%	31.18%
macroaverage		62.00%	55.84%	57.25%

Table 6.3: Individual prepositions results - L2 correct data

tell us if the model is performing in the same way on the L2 data as it does on the L1, or if different issues arise. The results of this analysis will help highlight areas of the model which need improving, and may raise important points relevant to the wider discussion on the application of NLP tools to learner language.

Table 6.3 gives the distributions of the prepositions within this set together with their precision and recall scores. The analogous table for the L1 data, Table 4.11, is reproduced here as Table 6.2 for ease of comparison. Overall, the distribution and relative frequencies of the data are very similar; a notable exception is that, as noted above, in the L2 text the most frequent preposition is ***in*** rather than ***of***, which is in fact also less frequent than ***to***. The learners' apparent overuse of ***in*** compared to the L1 data is a somewhat surprising finding; it can be surmised that this is related to particular constructions that appear more often in the type of texts produced by them, an hypothesis that will be investigated below. Generally, while one might expect significant differences in the distribution of content words in L2 vs. L1 data,

Target prep	Confused with								
	at	**by**	**for**	**from**	**in**	**of**	**on**	**to**	**with**
at	xx	4.65%	10.82%	2.95%	36.83%	19.46%	9.17%	10.28%	5.85%
by	6.54%	xx	8.50%	2.58%	41.38%	19.44%	5.41%	10.04%	6.10%
for	8.19%	3.93%	xx	1.91%	25.67%	36.12%	5.60%	11.29%	7.28%
from	6.19%	4.14%	6.72%	xx	26.98%	26.74%	7.70%	16.45%	5.07%
in	7.16%	9.28%	10.68%	3.01%	xx	43.40%	10.92%	8.96%	6.59%
of	3.95%	2.00%	18.81%	3.36%	40.21%	xx	9.46%	14.77%	7.43%
on	5.49%	3.85%	8.66%	2.29%	32.88%	27.92%	xx	12.20%	6.71%
to	9.77%	3.82%	11.49%	3.71%	24.86%	27.95%	9.43%	xx	8.95%
with	3.66%	4.43%	12.06%	2.24%	28.08%	26.63%	6.81%	16.10%	xx

Table 6.4: Confusion matrix for prepositions - L1 data

Target prep	Confused with								
	at	**by**	**for**	**from**	**in**	**of**	**on**	**to**	**with**
at	xx	3.62%	10.87%	2.17%	47.83%	15.94%	7.25%	5.80%	6.52%
by	16.35%	xx	6.73%	0.00%	49.04%	8.65%	8.65%	10.58%	0.00%
for	9.33%	5.22%	xx	1.49%	29.85%	35.07%	6.34%	7.09%	5.60%
from	15.89%	4.64%	8.61%	xx	33.77%	17.22%	7.95%	9.27%	2.65%
in	16.91%	7.35%	19.85%	2.21%	xx	35.54%	9.07%	5.64%	3.43%
of	3.88%	2.91%	13.59%	0.97%	54.37%	xx	6.80%	12.62%	4.85%
on	9.63%	2.67%	4.81%	1.07%	54.55%	13.90%	xx	7.49%	5.88%
to	25.20%	3.94%	8.27%	3.15%	27.56%	22.44%	6.30%	xx	3.15%
with	2.89%	6.36%	12.14%	1.73%	35.26%	25.43%	8.67%	7.51%	xx

Table 6.5: Confusion matrix for prepositions - correct L2 data

it would be unusual to find such differences in function words, as they are required at all levels of complexity.

An examination of the precision and recall scores raises some interesting points. Although the averages for both measures are within less than 2% of their L1 counterparts, which is to be expected given the minimal divergence in overall accuracy between the two models, these are not derived from similar sets of figures. Indeed, just over half the prepositions have different recall scores in the L2 data: the ones for ***by***, ***from***, and ***with*** are lower by 10% or more, while ***for*** and ***in*** have scores which are up to 14% higher. Several prepositions also display a drop in precision when compared to the L1 data. More specifically, ***at***, ***by***, and ***of*** have scores which are 7-18% lower, while an increase of around 10% is found for ***for*** and ***in***. A full explanation for these discrepancies can only be had by looking more closely at the data to detect whether there are any particular contexts which are causing the incorrect class assignations.

To assist in this analysis, a confusion matrix for Dapper's errors on this data is

presented in Table 6.5. It is analogous to the one for the L1 data, Table 4.14, which is also reproduced here for convenience as Table 6.4. It can be seen that there are many similarities between the two tables, but also several crucial points of divergence. The discussion that follows will focus on the latter, as it is believed that it is from these that potentially determining differences between the two text types can be identified. As a general point, however, it must be noted that it is not unusual to find some variation when there is such a difference in size between the two sets. Although the amount of data underpinning this analysis may not always be sufficient to warrant statistically reliable conclusions, overall it should be possible to discern some general trends of interest.

At

In Chapter 4, it is suggested that the pair ***at/in*** could be susceptible to confusion, as both items are used to indicate a static spatial relation. Here their reciprocal confusion is found to be about 10% higher than that observed in the L1 data. The PPs with ***at*** which are incorrectly assigned to the ***in*** class are mostly of two kinds: temporal expressions of the form *at first*, *at present*, *at 12pm/5am/etc.*, *at 65/18/other ages*, and locatives such as *at school*, *at this event*, *at the concert*, and so on, with the former type being more frequent. It is not clear why DAPPER finds such a strong relation between these temporal lexical items and ***in***[4]. This may point to something which requires more attention within the model itself, for example in its treatment of numerals. It is important to note that this type of confusion is not peculiar to the L2 data, as analogous misclassifications are found in the L1 results, too. However, phrases such as *at first*, *at present*, and *at 18* are particularly typical of student essays, where the learner is often required to produce a piece of (usually autobiographical) narrative regarding an event: these phrases serve as convenient markers to give the text a structural and chronological order. It is therefore likely that here they occur with a higher frequency than in the texts in the BNC, which in turn would lead to this particular type of confusion being observed more often.

The converse of this misclassification, that is, correct instances of ***in*** incorrectly labelled as ***at***, also regards mainly temporal expressions, especially those involving one of four particular lexical items: *time* (*in the time of/that...*), *moment* (*in this/that moment*), *end* (*in the end*), and *stage* (*in the last/first stages*). It is immediately evident that these are words which do occur often with ***at***, so it is clear that the classifier

[4]Nor is this the case for other similar lexical items: with words such as *moment* or *time*, there is a clear link with the preposition ***at***.

has identified a particularly strong relationship between them and that preposition. The issue of their confusability becomes more noticeable in this task because these phrases, too, are typical of the 'lexical toolkit' of the non-native writer and of the formal (but not necessarily academic) text he or she is expected to produce. What is beginning to emerge, then, is the finding that text type differences do have an impact on performance. This is due not so much to the presence of errors, as to the fact that the classifier appears less attuned to the more informal kind of language used by students, and typical of the narrative/expository texts that make up a large part of the L2 corpus.

On the other hand, it could be argued that several of the class assignations made by DAPPER for the incorrectly labelled instances of ***in*** are not in fact erroneous, and could perhaps even contribute to making the text sound more idiomatic. Indeed, although it is grammatical to have a sentence such as *in this moment, you have almost forgot the story you were reading* (or any of the other phrases mentioned above), using ***at*** *this moment* instead is also correct, and is more in line with native speaker intuitions. These types of errors on the part of the classifier, then, should not be seen as a failure to fully grasp the patterns of preposition occurrence in English, but as evidence of its having acquired the statistical tendencies of a particular type of writing, which unfortunately may not always be congruous to the students' essays. In this respect, using corrected CLC data or following the Chodorow et al. (2007) approach of including data from school textbooks in training may be beneficial as this may be more similar to the types of texts given to the system for testing.

By

A related confusion pair is the 10% increase in the number of instances of ***by*** incorrectly labelled as ***at***. These regard almost exclusively three types of phrases: *by the time that...*, *by the end*, and *by the age of...*. Similar considerations as above apply to the nature of these words and their frequent use in learner writing, easily explaining the reason for the confusion.

There is also an 8% increase in instances of ***by*** classified as ***in***. Just over a third of these consist of the phrase *by the way*, which the classifier assumes should be ***in*** *the way* instead. Presumably – and not surprisingly – this somewhat informal phrase does not figure very often in the training data, so the model is unable to recognise a strong relationship between these items. Conversely, it is observed that it is a very frequent phrase in the learners' essays, used especially in those which are intended to be letters to friends or colleagues. In the course of this analysis, then, it is becoming

increasingly clear that a relatively small number of semi-fixed phrases is at the root of the classifier's errors; this finding also lends further support to the claims that L2 writers tend to rely on a small set of lexical 'chunks', and to overuse them or display less variation when compared to L1 writers (see e.g. Granger 2004:134-136).

The other type of ***by***-PP often misclassified as ***in*** is that of the form ***by***+*-ing verb*, such as *by joining the club*, *by visiting his mother*, and so on. This is not as clear cut a case as the others discussed so far, as these phrases are indeed also grammatical with ***in***: cf. *by/in visiting his mother, he is doing an act of charity.* Arguably there are subtle differences in meaning, whereby ***in*** implies only an action that is taking place, while ***by*** can also be used for hypothetical cases: cf. ***by*** *visiting England, we could meet English people* vs. *?**in** visiting England, we could meet English people*[5] (but *we **have met** English people* is fine for both). This is also a misclassification which should not be considered a true error on DAPPER's part, and which would not result in a serious error should the suggested alternative be accepted.

From

Occurrences of ***from*** are, among other things, misclassified as ***at*** 9.5% more often than in the L1 data. Once again, three fixed phrases account for almost all these errors. Two of them are temporal expressions: *from the moment that...* and *from time to time*. The strong link between these lexical items and the prepositon ***at*** has already been observed. These phrases can also be considered typical of learner essays, and yet again show that the lexical item feature seems to override all other contextual characteristics. This becomes even more evident when the third phrase in this group is considered, namely *from my point of view*. The word *point* seems to trigger the label ***at***, perhaps on the model of *at this point*. The phrase *from my point of view* has a particularly high frequency in the essays, as students are often required to state their opinion on an issue or argue for or against a position; conversely, it is unlikely to have figured as often in the BNC. Although statements of opinions do of course appear there, too, they are more likely to be phrased with more variation and less reliance on this particular lexical chunk.

There is also a 7% increase in the number of instances of ***from*** classified as ***in***. Here, however, it is less easy to identify a clear pattern for the confusion. Several dates (e.g. *from 1995*, *from June*) are found, as are several common nouns denoting places (*I come from the city*, *from the countryside*, *from a place...*). While it is not unusual for these phrases to trigger the choice of the preposition ***in***, these are rather

[5] A ? indicates a sentence which is grammatical but does not sound very natural.

general phrases, not peculiar to learner writing, unlike the others discussed so far. It can be hypothesised that because of the lower frequency of ***from*** in the training data, these patterns have not been acquired as strongly as those for the far more frequent ***in***, and this is affecting the classifier's decisions for these phrases.

In

As well as the confusion with ***at***, correct instances of ***in*** also see an increase in the frequency with which they are mislabelled as ***for***. This occurs almost exclusively in temporal phrases, mostly with *year* (***in*** *the [previous] years* corrected to ***for*** *the [previous] years*) and *week* (***in*** *that/the past week* corrected to ***for*** *the past week*) as well as related items such as *minute*, *month*, and so on. These are not only further examples of semi-fixed phrases with high frequency in learner writing, but also another instance of the classifier making class assignations which do not result in ungrammaticality or, crucially, always in a change of meaning. Indeed, while in some cases the changed preposition would also change the meaning of the sentence entirely – cf. *I will be here* ***in*** *two weeks* vs. *I will be here* ***for*** *two weeks* – in many others both phrases are roughly equivalent – cf. ***in/for*** *the last few years, habits have been changing*[6]. This is a type of unnecessary correction which could only be avoided by including real-world knowledge to the task; its higher frequency in this dataset is due to the fact, already observed, that temporal expressions are particularly frequent in L2 writing.

Of

The number of times ***of*** is misclassified as ***in*** increases by 14%. This is a surprising finding, since in the L1 task this preposition tended to score very well due to, it was surmised, its high frequency in the data and fairly distinctive occurrence patterns. In fact, it is hard to discern particular contexts which trigger this error, although we recall that in the L1 data, too, ***in*** is given most often as the incorrect choice for ***of***. Two items stand out – phrases with locations as the object (*of the city*, *of the country*, etc.), and the phrase *of course*. The former can be explained by noting that overall it is more likely that those nouns are found as the object of ***in*** in the training data, leading to this incorrect choice. The latter, though – a phrase which accounts for over 20% of the incorrectly labelled instances of ***of*** – is more surprising. At a first

[6] A native speaker however observes that there is an implied difference between the two sentences: in the first case, the action described is understood to be sporadic, while in the second, it is understood to be continous over the period of time specified – cf. also an example such as ***in/for*** *the past year, I have been baking cakes.*

analysis it would not appear to be a structural or chronological marker as so many of the other phrases discussed, and therefore not particular to learner writing. However, upon reflection it emerges that its relative infrequency in the BNC compared to other phrases where the preposition involved is ***in***, which must be the cause of the error, is not so strange. The phrase *of course* belongs to an academic style of argumentation that is more informal and typical of an insecure writer than what one would find in the BNC, and as such, figures very often in student essays if the writer is not fully confident of the force of his or her argument. Although not a structural marker as such, it is nevertheless a hallmark of student writing, analogous to the others identified so far.

On

Another preposition to show a sharp increase (22%) in its confusion with ***in*** is ***on***. In the discussion of the L1 data, it had been hypothesised that these two prepositions could have been confused with each other often, especially in the spatial domain. However, that did not occur as frequently as expected, nor is it the source of errors here. In fact, all of the contexts triggering errors are either temporal or idiomatic. For example, there are phrases such as ***on*** *that/the/this day...* or ***on*** + a date (*December 15*, *May 1*, etc.). It is not clear why DAPPER would mislabel these instances, as the resulting phrases would not be grammatical (e.g. *****in*** *that day*) nor would one expect such phrases to be more common in the BNC than the correct ones. An example of idiomatic use that stands out, because of its implications rather than its frequency, is ***on*** *the internet*. This lexical item is barely found in the BNC (4 times only) so it is not surprising that the model does not have a clear picture of its preferred co-occurrences. A further potential pitfall is therefore introduced, along with differences in style and structure of the text: the use of individual lexical items which are peculiar to a contemporary student's life or work, and highly unlikely to figure prominently in the BNC because of the time at which the latter corpus was created.

By far the greatest source of misclassifications – around 50% – is the phrase *on the one/other hand*. This, together with *from my point of view*, is one of the most typical phrases to be found in argumentation essays, which are the backbone of L2 writing assignments. Its high frequency, especially compared to the texts in the BNC, suggests that learners are producing texts which are structured around 'prefabricated' chunks (Granger 1998c), and should be encouraged to explore other means of expressing opinions and facts in their writing.

To

The most marked change in errors involving occurrences of ***to*** is the fact that they are misclassified as ***at*** with an increased frequency of 15% compared to the L1 data. There are some phrases which recur relatively frequently in this group; they include expressions such as *looking forward* ***to*** (where the verb *look* evidently triggers ***at***) and *increase/decrease* ***to*** *the level of...* or ***to*** *[some digit] pounds* (e.g. *the price increased* ***to*** *1500 pounds*). In the latter examples, it is assumed the object nouns are driving the choice of ***at*** (cf. *at this level*). It is possible that these expressions figure somewhat more frequently in learner essays than in the BNC because of the inclusion of Business English exams in the corpus, where one of the tasks usually requires the writing or interpretation of business charts and other similar data.

There is also another major group of phrases which is susceptible to this misclassification. This is linked to the issue of the inclusion of ***to*** already addressed in Section 3.3.1, namely the fact that because this preposition is tagged as 'TO' whether it is a preposition or an infinitival marker, the model is trained to recognise both of these uses. Here this becomes particularly evident: almost half of the instances of ***to*** mislabelled as ***at*** are sentence-initial phrases where ***to*** appears as an infinitival marker, such as: *to start with...*; *to attract more customers...*; *to conclude...*, and so on. These phrases, together with similar verb-less ones such as *to this end*, appear with high frequency in sentence-initial position in learner writing, again, it is assumed, as a useful way to mark the logical ordering of the text. However, they are much less frequent in the L1 data where, instead, adjuncts headed by ***at*** can be often found in the same position. It is plausible that despite the fact that the use of ***at*** in these contexts results in ungrammatical expressions, this preposition is being chosen in virtue of its being much more likely than ***to*** to occur in sentence-initial position.

With

Finally, as regards ***with***, not many changes in its confusion patterns are observed, apart from a 7% increase in the number of times its instances are labelled as ***in***. An analysis of the incorrectly labelled cases reveals a peculiarity of L2 writing. It emerges that learners often choose to begin their sentences with ***with***-PPs of a form which can be compared to the Latin ablative absolute in function, as can be seen by the following examples:

(8) With her mind panicking, she called the police.

(9) With his voice strong and confident, he began to speak.

These phrases are not as a rule marked by the CLC annotators as incorrect, although they clearly sound stilted and unnatural to many fluent speakers of English; it can only be assumed that they are the result of interference from the L1[7]. The preposition suggested by DAPPER for these cases is ***in***; while this does not improve the readability of the sentence, it points to the underlying fact that English sentences rarely begin with the preposition ***with***. They do, however, often begin with adjuncts headed by ***in***, so it is possible that the classifer is drawing on this knowledge to make its (incorrect) class assignation.

In the discussion above, those pairs where confusability is lower than in the L1 data have not been addressed, although these do occur. It is assumed that their occurrence is also an effect of differences in text type: the elements that make certain pairs of confusables more likely in the BNC data may not be as frequent in the L2 data.

6.3.1.1 Preliminary conclusions

The previous section discussed a number of features of L2 writing which are believed to be the cause of errors for the classifier. These have proved to be partly different from what anticipated in Section 6.1. It seems clear that misspellings and other ill-formed input are not the main source of difficulty for DAPPER, although it could be argued that text where prepositions are used correctly is also more likely to be generally well-formed. However, errors do occur, and are often found in instances which the classifier has assigned to the wrong class. Examples of misspelled words include: *by all meance* (*means*), *viewus of* (*views*), *with enthousiasm*. In all these cases, the preposition chosen by the student is correct, and indeed has a fairly strong collocational link to the lexical item; but DAPPER does not recognise the misspelled item and selects a different, incorrect preposition instead. This is the kind of problem that can be easily solved by introducing a spellcheck filter at the preprocessing stage.

Grammatical errors also sometimes lead to incorrect class assignment, as in the following example:

(10) On top of that this amount will be **increase** [**increased**] **with** premiums for special wins.

[7] In fact, both of these examples were produced by learners with the same L1.

Here, the verb *increase* is in the wrong form. This leads the parser to tag it as a noun rather than a verb, in turn triggering the choice of the preposition ***in*** for that particular context, which is not surprising given that *an increase* ***in***... is a fairly common phrase. However, it is wrong in this case, as the original ***with*** is not incorrect. This is an example of a well-formed sentence where a single grammatical error is sufficient to lead the classifier astray. Unfortunately, these errors are also hard to detect or filter out without relying on the annotators' tagging, which of course would not be present in instances not taken from the corpus.

The main factor affecting DAPPER's performance is, as anticipated in the previous section, the different syntactic structure used by learners. This is represented both by a higher-than-average use of particular phrases and discourse markers, and by a tendency to place any kind adjunct clauses, not just temporal and locative ones, in sentence-initial position (a characteristic also documented e.g. in Granger (2004:135)). While not ungrammatical, these rhetorical strategies are perhaps not always ones that L1 English writers would use, and are especially typical of text types which are not very frequent in the BNC, such as student essays.

Making the model more familiar with different kinds of adjuncts in initial position could be achieved by including different types of texts at the training stage, for example ones more similar to the essays written by students. This would also have the advantage of potentially including more of the fixed phrases that have been found to be so favoured by learners. Another possible solution to the issue of fixed phrases being mislabelled would be to enhance the model so that it recognises a particular set of semi-fixed expressions and always defaults to the correct preposition for those. This could be done fairly easily, since the set of such expressions is not very large.

A different type of solution, and one which will be discussed more fully in Section 7.2, is to allow the classifier to suggest more than one choice and rank them in order of how confident it is about each. An informed user could then review this list and choose the preposition most appropriate for his or her intended meaning. This solution, however, would only be of real assistance to a more advanced or confident student.

In light of the analysis above, the trends in precision and recall observed no longer appear so surprising, as they are driven by the peculiarity of the texts used to assess DAPPER. From a pedagogical perspective, the claim that L2 writers tend to rely on a small set of fixed expressions is reinforced. It appears that the texts produced by these students, though generally not ungrammatical, are nevertheless rather different from standard English texts. Students should be perhaps encouraged to try a variety

of approaches and syntactic structures, if their final aim is to sound as native-like and fluent as possible.

6.3.2 Performance on incorrect data

To test DAPPER's ability to identify and correct preposition errors, the set of 1116 instances of erroneously used prepositions was submitted to it. Of these, the system flagged the presence of an error – i.e. found a disagreement between the preposition it believed to be most appropriate for that context, and the one used by the writer – in 76.43% of cases. At first glance, such a high accuracy score is encouraging. However, it is not enough to be able to point to a perceived error in the text; the system can be considered successful only if this is accompanied by an appropriate suggestion for an alternative preposition choice.

All 1116 instances used in this task have been manually analysed to attempt to gain a clearer picture of how DAPPER is performing and what kinds of problems it may be encountering. As mentioned above, over 76% of instances are marked as containing an error. However, of the ones flagged as incorrect, just over half (51.70%) are also corrected appropriately. This means that of 1116 errors, the proportion identified and corrected is a mere **39.51%**.

While it is encouraging that in principle the system developed can indeed be used to recognise and correct preposition errors, such a low recall score is not ideal, as it would still leave far too many errors undetected, potentially misleading the student about his or her real level of achievement. However, results reported on similar tasks in the literature suggest that this is a challenging task. Tetreault and Chodorow (2008b), for example, achieve up to 84% precision and 19% recall; as will be shown below (Section 6.3.2.1), DAPPER achieves a higher recall, but at the expense of much lower precision.

In trying to understand the possible reasons for the gap between performance on correct text (whether L1 or L2) and incorrect L2 text, the incorrectly labelled instances are inspected with particular regard to three factors, which until now have not been identified as playing a significant role: interrelation with another type of error[8]; disagreement with the annotators; grammatical acceptability.

Interrelation with other types of errors refers to those cases where the preposition error is due to the presence of another error in its immediate context. As seen in Chapter 2, this is an issue which is often remarked on in work describing error

[8]These are referred to as "concomitant errors" by (James 1998:116): those which occur because an erroneous form was selected somewhere previously in the sentence.

correction of particular kinds of errors. Concomitant errors include not only spelling mistakes and incorrect POS, as discussed in the previous section, but also more complex cases where the error lies in the choice of lexical item as head or object of the preposition. An example is the sentence below, shown with and without its corrections[9]:

(11) He greeted me for a lunch there and I greeted him for a drink.
He **greeted** [**treated**] me **for** [**to**] **#UD a /#UD** lunch there and I **greeted** [**treated**] him **for** [**to**] a drink.

For these preposition contexts, the classifier chooses the class ***with***, presumably on the basis of the high frequency of phrases such as *greet with a smile, a kiss*, and so on. However, the problem in this sentence clearly lies in the choice of verb rather than preposition, and indeed the annotators mark this by suggesting a different verb. The correct verb requires a different preposition, which leads to the preposition in the text also being marked as erroneous – but Dapper of course does not have access to the corrected version of the text, as all it sees is the sentence before any annotations have been marked. It is therefore unlikely in these cases that the preposition suggested is the one noted as appropriate by the annotators. Arguably, this is an issue related mainly to the annotation scheme: it might be more appropriate to tag these cases as a single error rather than two separate ones, as the error effectively consists of a wrong lexical item rather than a wrong lexical item and a wrong preposition.

Despite the perceived high frequency of spelling errors and misused lexical items, and the expectation that they would prove a significant obstacle to good performance, this analysis reveals that these factors account for only 3% of cases where Dapper did not assign the correct label to an instance. The presence of other errors within the PP is not, therefore, the main cause of problems for the system.

The other two factors mentioned above, disagreement with the annotators and grammatical acceptability, are best treated together. Both refer to cases where the classifier selects a preposition which is correct in the given context, but is not the correct one in that particular case, either because it is not a pragmatically[10] appropriate choice, or because, despite being pragmatically and grammatically appropriate, it is

[9]In the presentation of learner sentences, both the original version and the one with the annotators' corrections (in square brackets) will be given if required for ease of readability; in some cases, for reasons of clarity, annotations not relevant to the discussion will be omitted.

[10]The term 'pragmatically' is used loosely to refer to the discourse or contextual aspect of the sentence, to distinguish cases where the selected term is incorrect because it is not licensed by the grammar from those where the properties of the context make it implausible or inappropriate.

not the preposition selected by the annotators for that particular context. The former case includes examples such as the following:

(12) It #TV has [had] /#TV been left **#RT in /#RT** the bus.
Annotators' decision: **ON**
Classifier decision: **BY**

(13) The view **#RT in /#RT** Interlaken is #DJ wonderfull [wonderful] /#DJ.
Annotators' decision: **FROM**
Classifier decision: **OF**

In these examples, the preposition suggested by DAPPER is correct and yields a grammatical sentence, albeit one with a rather different meaning – a problem also encountered in the analysis of L1 results. These errors are evidence of the system using the linguistic knowledge it has acquired to inform its choices, and as such, constitute a different type of error, one which could be addressed, as already mentioned, by allowing the classifier to suggest more than one option, assuming such an output would contain the more appropriate preposition, too. Classifier errors stemming from a grammatically correct but pragmatically incorrect choice account for around 7% of all errors. Despite being a relatively large amount, it is impossible to prevent these errors altogether without knowledge of the wider discourse; the best approach is therefore to find a solution that deals with their occurrence, such as ranking more than one suggestion.

In the discussion of L1 results, it was noted (Section 4.1) that the findings highlighted by the human judgement task showed that it is often possible for the same context to license more than one preposition without necessarily bringing a change in meaning. This is even more evident with regard to examples of class assignments by the system where its choice is correct in terms of both meaning and grammar, but happens not to be the preposition found in the text against which the classifier's performance is assessed. In the specific case of L2 preposition errors, this refers to cases where the classifier's suggested correction for an error it has identified does not correspond to the one proposed by the annotators. Since their corrections are used as the benchmark against which DAPPER's performance is evaluated, these instances are counted as the classifier being wrong. Some examples of this are:

(14) I've known him since we were **#RT on /#RT** primary school.
Annotators' decision: **IN**
Classifier decision: **AT**

(15) The student also can misuse the #CN informations [information] /#CN given **#RT in /#RT** the computer.
Classifier decision: **ON**
Classifier decision: **BY**

As in the previous examples, the classifier's choices here are grammatically correct; additionally, they also yield essentially the same meaning as the prepositions selected by the annotators (arguably in the second case ***by*** is more idiomatic than ***on***). The difference in choice lies in the individual preferences of the annotators, which are by no means consistent: for example, there are instances of sentences such as *I live* ***at*** *Green Street* corrected as both ***on*** *Green Street* and ***in*** *Green Street*, with no apparently discernible pattern to their choices[11].

Examples of this sort should not be considered errors on DAPPER's part, as the system fulfils its brief by correctly signalling the presence of an error and offering a correction for it which is well-suited to the context, regardless of its agreement with the annotators. If these cases are excluded from the error count, a more realistic indication of the model's performance is obtained, one that is free from the annotators' bias. These 'non-errors' account for 11% of the classifier's reported errors, making disagreement with the text the most prominent among the factors involved in affecting its performance.

On the basis of this analysis, the classifier's accuracy score can be recalculated to include such correctly identified errors. It is assumed that of the instances initially marked as mistakes made by the classifier – either because the error wasn't spotted or because the suggested correction was not the same as the one chosen by the annotators – 11% should be considered as correct instead. This brings the accuracy to around 46%. It could also be claimed that the true measure of DAPPER's error rate on this task should also exclude the two other types of errors discussed above, since those effectively relate to issues other than those of error identification and correction.

[11] As there are five different annotators, these divergences are to be expected; however, there is no way of telling which texts have been annotated by each of them. It has been suggested to me that these preferences may lie in differences between American and British English, but all annotators are British English speakers.

Furthermore, these represent the first results achieved by DAPPER, without further filtering of the input data or the use of additional components such as language models. Differences in the task setup, as already mentioned, make it difficult to make direct comparisons with other recent work in the area: however, it is clear that this it is a task not easily solved. Gamon et al. (2008), for example, report a score of 56%, although this refers to all three types of preposition error, namely omission, redundancy, and substitution.

6.3.2.1 Precision and recall on erroneous instances

So far the extent of the classifier's accuracy in identifying errors has been discussed only through a quantitative rather than a qualitative overview of the results. In this section, the results will be analysed in greater detail, to gain a better understanding of where the system may be failing. The original results will be referred to, taking as incorrect all those where there is disagreement with the annotators[12].

The first issue to address is whether the distribution of the erroneous prepositions in the text mirrors that of their correct counterparts. This would not only give useful insights into the error patterns of L2 writers, but can also help explain some problems in DAPPER's performance, if the texts are found to present a different distribution from that observed in training. Indeed, there are several significant divergences, and it is not always the case that prepositions which are most frequent in L1 English are also the ones most frequently misused by learners: notable examples include ***at***, which occurs in erroneous constructions with a much higher frequency than its observed frequency in the language would lead one to expect, and ***to***, which displays the opposite behaviour, being tagged as incorrect much less frequently.

While closer analysis of these figures would be of great assistance in sketching a picture of patterns of L2 preposition use and acquisition, the focus in this discussion is on the performance of DAPPER rather than on that of the learners. The issue to be considered therefore is the ease with which the system was able to signal and correct errors involving the various prepositions; precision and recall figures will supply this information. Similarly to the issue of 'learnability' addressed in Section 4.3, the discussion here will centre around 'correctability': examining whether any misused prepositions are left uncorrected more often than others, which is a reflection of the reliability of the models acquired both for the misused preposition and for the

[12]Due to intellectual property restrictions, the discussion in this section cannot always report precise figures and statistics. For example, precise figures for the occurrence of particular prepositions as errors cannot be given; only the general trends observed can be described.

	Proportion of training data	Precision	Recall	F-score
of	27.83% (2,501,327)	74.28%	90.47%	81.58%
to	20.64% (1,855,304)	85.99%	81.73%	83.81%
in	17.68% (1,589,718)	60.15%	67.60%	63.66%
for	8.01% (720,369)	55.47%	43.78%	48.94%
on	6.54% (587,871)	58.52%	45.81%	51.39%
with	6.03% (541,696)	58.13%	46.33%	51.39%
at	4.72% (424,539)	57.44%	52.12%	54.65%
by	4.69% (421,430)	63.83%	56.51%	59.95%
from	3.86% (347,105)	59.20%	32.07%	41.60%
macroaverage		63.67%	57.38%	59.68%

Table 6.6: Individual prepositions results - L1 data

	Precision	Recall	F-score
of	21.29%	63.10%	31.83%
to	49.22%	33.87%	40.13%
in	34.72%	59.69%	43.90%
for	48.46%	39.87%	43.75%
on	70.09%	35.05%	46.73%
with	18.75%	11.54%	14.29%
at	48.96%	35.34%	41.05%
by	33.33%	22.22%	26.67%
from	52.94%	16.67%	25.35%
macroaverage	41.97%	35.26%	38.18%

Table 6.7: Individual prepositions results - incorrect L2 data. Exact frequency figures cannot be given due to intellectual property restrictions.

intended correction. In other words, given an erroneous instance of ***at*** successfully corrected to ***in***, is this due to the model's good knowledge of uses of ***at***, of ***in***, or of a combination of the two?

Table 6.7 presents recall and precision scores for each preposition, together with the analogous figures from the L1 task, Table 4.11, repeated here as Table 6.6 for ease of comparison. Recall refers to the system's success at recognising that the target, correct, preposition is required in a given context. In other words, it measures its success at detecting misused prepositions: low recall means that too many errors are going undetected. It is calculated thus:

$$Recall = \frac{\#\ of\ times\ need\ for\ target\ prep.\ identified}{\#\ of\ times\ target\ prep.\ needed}$$

Precision shows what proportion of instances assigned to a particular class do belong

to that class. The formula is given below. Low precision means that the class labels being assigned are often inappropriate and that therefore errors are not being corrected appropriately.

$$Precision = \frac{\#\ of\ times\ class\ label\ assigned\ correctly}{\#\ of\ times\ class\ label\ assigned}$$

As a general consideration, it is noted that average recall and precision are lower than their L1 counterparts, which is expected given the overall lower accuracy obtained. Similarly to the L1 data, precision is about 10% higher than recall, which is positive because, as already observed, in this kind of application it is important to ensure precision is not too low. One can observe that the average scores for this task are about 20% lower than those obtained from the BNC data; this figure could be a measure of the gap that remains between the two tasks and the texts they involve. While it is not insignificant, it is not unsurmountable, and it is possible that with filters in place to account for the peculiarities of L2 data, this figure can be reduced.

In what follows, each preposition will be analysed individually, analogously to the previous section, to assess particular issues which arise with each and establish whether specific causes for low performance can be identified. This closer analysis is necessary for a full understanding of what does and does not work well in this task, as opposed to merely obtaining a general accuracy score, as it will allow more precise fine-tuning of the various components of DAPPER.

Looking at recall scores, it can be seen that all the prepositions, with the exclusion of ***for*** and ***in***, display a marked drop, and it is on these divergences that the analysis will focus. ***Of*** and ***in*** are still the best-performing prepositions, although not by such a large margin, but performance on ***to*** is much degraded, an issue which will be addressed in the discussion below. At the opposite end of the scale, performance on ***from*** is still very poor but is no longer the lowest: this is now ***with***. To assist in the analysis of recall scores, the concept of **correct frequency** is referred to. This is the frequency a preposition would have if it had been correctly used where expected at all times. These scores are compared to the distribution of each preposition in the BNC data, as differences between the two may help explain some of the problems encountered, as already noted in Section 6.3.1.

As for precision, it is obvious that, as the overall average is much lower than the L1 data, most individual scores will also be lower than their L1 counterparts. Indeed this occurs in every case except ***on***. These low precision scores are an alarming shortcoming of the system as they indicate that the learner would potentially receive

much incorrect and misleading information. It is important therefore to look at the types of phrases which are causing these errors, to ascertain whether the confusion patterns are analogous to those observed for the L1 and/or the correct L2 data, in which case the solutions proposed for them can also be applied here, or if there are additional issues that need to be addressed. In this discussion, the focus is only on those prepositions which display a larger change in precision.

At

Recall on ***at*** is lower by 17%. Its correct frequency is somewhat higher than the distribution observed for it in the BNC, which may help account for low recall: there must be a wider, or different, range of ***at***-PPs being used by learners which are not so prominent in training and therefore not acquired as strongly. Instances of ***at*** as the target preposition go mostly uncorrected because the classifier believes they should be assigned to ***in*** or ***of*** instead. So, if the student has used one of those two prepositions, the error goes undetected; if he or she has used another, also incorrect preposition, the classifier notes the error but inappropriately suggests ***in*** or ***of*** as the correction. A large proportion of these mistakes are found to involve a small number of contexts, as already noted in Section 6.3. Among these is the presence of digits in the object position, for example expressing percentages or times: cf. *it start* ***from*** *6:00 pm*, or *although USA had the highest productivity per worker (140) hourly output was the second lowest* ***of*** *approximately 105*, both examples of cases where the preposition instance is assigned to the class ***in***. It has already been noted both that the system appears to struggle with numbers, and that they are found to occur particularly frequently in L2 writing tasks, suggesting this is an area on which further improvements must be focused. There is also some confusion with locatives, especially when the object is a word such as *house* (expected correct phrase: *stay* ***at*** *my house*), *office* (*a meeting* ***at*** *the office*), or *company* (*the staff* ***at*** *the company*). Again, these are terms for which the classifier has acquired a particularly strong collocation with ***in***, which is not incorrect; however, the fact that they occur frequently in these texts with a preposition with which they have a weaker association is a major factor in negatively affecting recall on this preposition.

By

In the L1 analysis the relatively strong performance of this preposition was remarked upon, but here low recall and precision scores are observed, with a difference of 30% or more. However, it must also be said that there is only a small number of

instances where ***by*** is the target preposition, which makes it hard to draw any reliable conclusions. On the other hand, its correct frequency is not dissimilar from the one found for it in the L1 data, so this does not point to any unusual usage patterns on the part of the learners which could be playing a role in recall. Equally, this preposition is also chosen rarely (24 times only) as a class label in this subset of the data, a paucity of instances which makes it difficult to come to any valid generalisations. It is observed, however, that the confusion with ***in***, observed in the correct L2 data, is also found here. While many of the classifier's failures to correct instances where ***by*** is the target preposition are not easily explained, several others belong to contexts where the preposition suggested by the classifier is equally acceptable, as in the following two examples:

(16) Do you can going **#RT with /#RT** the bus and train.
Annotators' decision: **BY**
Classifier decision: **ON**

(17) I'm very delighted **#RT for /#RT** your proposition.
Annotators' decision: **BY**
Classifier decision: **WITH**

In both examples, the preposition chosen by the learner is clearly wrong (and one could perhaps even identify L1 traits which could have triggered them). The suggestion made by the classifier, while not in agreement with that of the annotators, is nevertheless both grammatically and pragmatically plausible with little or no change in meaning. From a linguistic perspective, it is interesting to see how the instrumentality expressed by ***by*** can also be communicated by other, more common prepositions; this could perhaps also be of assistance in a pedagogical framework, in encouraging learners to use different prepositions to convey the same proposition if they appear to be more comfortable with ones they encounter more often.

For

As mentioned above, recall for this preposition does not show much variation between the two tasks, suggesting Dapper is not encountering any additional problems to those already discussed in Section 4.3 regarding its recall in L1. It is noted, however, that its correct frequency is somewhat higher than in the L1 data (this is also

the case in the correct L2 data, where its frequency is higher by 3%); this may be one of the very few instances where a difference in distribution does not negatively affect performance.

From

Recall on ***from*** is here even lower than the already rather low score achieved in the L1 task. Although its frequency is similar to the L1 data, it does not occur often as a target preposition so, again, reliable conclusions are hard to draw, and it is particularly hard to discern a typical pattern in the classifier's mistakes. In many cases, the preposition given by the classifier instead of the correct ***from*** is also grammatical, but is not always appropriate in the context – cf. the following examples:

(18) The turnover **#RT by /#RT** heaters was constantly decreasing.
Annotators' decision: **FROM**
Classifier decision: **OF**

(19) It was my cousin **#RT of /#RT** Madrid.
Annotators' decision: **FROM**
Classifier decision: **IN**

(20) You know that the ring was very important to me, because it's a present **#RT of /#RT** you!
Annotators' decision: **FROM**
Classifier decision: **FOR**

It is probable that the low frequency of this preposition in the language makes the classifier less likely to select it as a class label, especially when there are other more frequent prepositions which could fit the context. In the last example, in particular, it is impossible to know which preposition is correct between ***for*** and ***from*** without knowledge of the wider discourse. This is a problem that Dapper is currently not able to solve (whereas the issue of antonymical prepositions, for example, is addressed by Tetreault and Chodorow (2008b)), though the option of outputting more than one suggestion may go some way towards overcoming it.

In

As noted above, recall figures for this preposition do not decrease too much compared to the L1 ones; its correct distribution in this data is very similar to that displayed

in the L1 text, which may be one of the reasons for the relatively small gap in performance. It is assumed that the kinds of problems causing a low recall are similar to those already discussed in other sections of this chapter.

On the other hand, precision is much lower, which is not surprising if we note that it is given as a class label in almost a third of all cases. A general tendency in DAPPER to choose ***in*** as a class label most often has been observed before, but in this task this is even more evident. An overview of the data reveals that the problems affecting precision of this preposition are similar to those found in the L2 correct data task, namely centred mostly around temporal expressions, especially specific dates. For example:

(21) The training programme will start **#RT at /#RT** the 1st August 1999.
Annotators' decision: **ON**
Classifier decision: **IN**

This kind of error on the part of the classifier has already been discussed along with other kinds of temporal expressions and observations on the increased frequency of simple temporal and locative expressions in L2 essays. These types of expressions, especially dates, also appear to be a major source of confusion for learners and therefore their presence is particularly high in the erroneous instances dataset, a fact which is not insignificant in accounting for the drop in precision. The confusion between ***at***, ***in***, and ***on*** is also exemplified by another type of phrase which appears often in the data: *call me* ***on*** *[some phone number]*. This phrase, which is very frequent due to the presence of Business English test data, tends to occur in learner writing with the incorrect preposition ***at***. The classifier, however, displays a strong preference for assigning the preposition ***in*** to these PPs, perhaps influenced by the numbers that follow. This, too, contributes noticeably to lower precision. This brief analysis of issues surrounding the use of ***in*** and other locative prepositions reinforces the findings that the type of text and vocabulary used by learners can be a significant factor in affecting the performance of a system trained on different data.

Of

Of stands out in the L1 results for achieving high precision and recall scores; here, the opposite occurs, namely low recall and precision (lower by over 30% and 50% respectively). This would appear to partly contradict earlier findings regarding the strength of the model for it acquired by the system. It is also interesting to observe that the correct frequency for this preposition in the L2 erroneous data is much lower

than the L1 frequency. This could mean that learners are comfortable in using it and rarely make mistakes in which it is involved. However, this hypothesis is not suppported by the observation that in the correct L2 data, too, its relative frequency is much lower than in the L1 data. Perhaps, then, learners prefer to avoid its use and rely on other constructions where the preposition does not occur, although given its pervasiveness in the language this may not always be easy. Whatever the cause, the skewed distribution may well play a role in negatively affecting both recall and precision, as the classifier has been trained to expect it to occur with a much higher frequency. Indeed, it is suggested often as the class label for a given instance, despite not being needed so frequently.

Particular contexts which may be the cause of low recall are harder to identify than for other prepositions. Several seem to belong to instances where the preposition selected by the classifier instead is also grammatical, but often carries a different meaning. Some examples are:

(22) We always celebrate the birthday **#RT for /#RT** our family's member [family members].
Annotators' decision: **OF**
Classifier decision: **WITH**

(23) The problem was that the software **#RT in /#RT** our company didn't work.
Annotators' decision: **OF**
Classifier decision: **FOR**

There are several peculiarities evident in these sentences, and in others like them in the data, which are likely to be at the root of the classifier's difficulties in assigning the instances to the correct class. For example, the structure of the sentences does not follow an order which sounds most natural in L1 English, consistent with a general tendency found in L2 writers to underuse the possessive marker in favour of **of**-constructions (which is understandable, as it may be more similar to their L1). As already remarked in discussing the L2 correct data, such differences in sentence organisation and choice of phrases often have a negative effect on the system's performance[13]. It is also interesting to note that often the prepositions being incorrectly suggested by Dapper are low-frequency ones: this shows that frequency is not the

[13]Arguably the sentences shown here would sound much more 'correct' if they were rewritten entirely to eliminate the use of the preposition ***of*** in this way, but this is not in line with the corpus annotation guidelines.

overriding criterion in preposition selection by the system. The strength of the relation with particular lexical items, combined with the additional challenge of unusual sentence structure, must be leading the classifier towards the incorrect choices.

As regards precision, one source of confusion seems to be those instances where ***for*** is correct instead, as already noted in Chapter 4 with regard to L1 confusable pairs. This surfaces in phrases which are marginally acceptable as correct English, leading DAPPER not to recognise the error. In both of the following examples, the correct preposition should be ***for***, not ***of*** as selected by both student and system:

(24) I have the brochure **of** the company.

(25) She drew the plans **of** the apartments.

Another issue relating to this preposition, which is unlikely to be specific to L2 writing, is the trigger constituted by several lexical items which do occur very often with ***of*** in the language, but not to the exclusion of other prepositions: words such as *view, think, quality, benefit, problem, course* to name but a few[14]. In the instances under consideration, these words appear with an incorrect preposition, but the correction suggested by DAPPER, ***of***, is equally wrong in this case, as the examples that follow show. This type of issue could be addressed by allowing the classifier to make and rank more than one suggestion.

(26) Courses **#RT for /#RT** health and safety laws
Annotators' decision: **ON**
Classifier decision: **OF** – influenced perhaps by *course of action, course of treatment*, etc.

(27) There is a nice view **#RT in /#RT** my window.
Annotators' decision: **FROM**
Classifier decision: **OF** – influenced by *a view of... a city, castle*, etc.

(28) This can be a #RJ bonus /#RJ quality **#RT to /#RT** a person.
Annotators' decision: **IN**
Classifier decision: **OF** – influenced perhaps by *quality of life, quality of this product*, etc.

[14]While these are not very unusual words, it could be argued they are somewhat more likely to occur in business and opinion-giving texts, thus reflecting the topics of the learner essays yet again.

On

The drop in recall for ***on*** is comparatively small, which is somewhat surprising given the fact that its correct frequency here is rather higher than in the L1 data. This discrepancy in distribution is probably due to the high frequency of PPs involving dates and phone numbers (*meeting* ***on*** *August 1, call me* ***on*** *123455*) which seems characteristic of the learner texts. The types of contexts where the need for this preposition was not correctly recognised are similar to those already discussed above. Usually the need for this preposition is not recognised, the instance being instead assigned the label ***in*** or ***of***. As the system has already been shown to have acquired an imperfect model for ***on***, it is not surprising that in data where it is even more frequent, recall will be lower.

Conversely, ***on*** is the only preposition to show an improvement in precision compared to the L1 data, of almost 12%. How can this isolated inversion in the trend be explained? The majority of instances presenting errors where ***on*** should be used rather than the preposition chosen by the learner involve the prepositions ***at*** and ***in***. As already remarked, these three prepositions are often confused by L2 writers, and confusion between ***in*** and ***on*** is also found in the L1 results. Indeed, not all such errors are recognised by DAPPER, along patterns already discussed; but several of them which involve simple, common words are. Typical examples of incorrect phrases produced by learners include ***in*** *the bus,* ***in/at*** *the [first, second] floor,* ***at/in*** *XYZ Street*. The high frequency of these lexical items in the language, and the strong collocation pattern acquired with ***on*** in training, ensures that these errors are dealt with appropriately.

As for cases of failed error recognition, i.e. not realising that ***on*** is the correct label for a particular instance, along with the confusion with the locative prepositions already mentioned there is a tendency for this to occur when the student has incorrectly used ***of*** and DAPPER does not register the presence of an error. This is not so evident in the L1 data, so it must be assumed that it is due to pecularities of L2 writing. While a strong pattern does not emerge, most of these cases appear to cluster around a few lexical items such as *report* (e.g. *I will write the report* ***of*** *the meeting*) or *ideas* (*we discussed some good ideas* ***of*** *staff development*). One would have thought that these words are not particularly characteristic of learner writing, but evidently their frequency is not as high in the BNC, hence the lack of strong collocational patterns for those combinations.

To

As anticipated above, ***to*** displays a very large decrease in recall which makes it one of the worst rather than best performing prepositions, unlike in the L1 data. Although its correct frequency is also somewhat lower than in the L1 data, this difference is not so large as to warrant such a low recall score. It is hard to pick out major error triggers for this preposition. Many of the classifier's mistakes do not seem to have a clear origin. For some, lexical item effects are certainly at work, such as the already mentioned *looking forward* ***to*** which is regularly corrected to *looking* ***at***. The confusion between ***to*** and ***for*** has been mentioned in conjunction with the analysis of ***for***, and these results confirm that finding. It would seem, from looking at the types of contexts where the correction is inappropriate, that while DAPPER is fairly confident and successful at dealing with the concrete, spatial sense of **to** (e.g. *going to, visit to*), metaphorical uses are harder to acquire, cf.:

(29) It was realy a good start **#RT for /#RT** a wonderfull holiday.
Annotators' decision: **TO**
Classifier decision: **OF**

(30) She's really a good friend **#RT for /#RT** me.
Annotators' decision: **TO**
Classifier decision: **OF**

(31) Reply **#RT for /#RT** the letter
Annotators' decision: **TO**
Classifier decision: **IN**

In all these examples, the preposition chosen by the classifier is ungrammatical in the given contexts. However, it may be also often found in the language in phrases including at least some of the lexical items surrounding it, which might explain the wrong choice (cf. *start* ***of****, friend* ***of****,* ***in*** *the letter*). This is a way of using ***to*** which is indeed less frequent and hard to give clear usage rules for, so it is not surprising that learners find it difficult to master, too.

Precision drops by over 35%, which is surprising given it is usually an easily identifiable preposition. Indeed, in several instances where a learner did not use it where it would have been correct, DAPPER successfully spots the error and gives the appropriate correction. This includes mainly errors in locative PPs such as *go in/at, a visit in, drive in*, and so on, where obviously ***to*** is the correct choice. Where the system is less successful is, as already observed, in distinguishing ***to*** and ***for*** when

they are being used in a benefactive sense. This is an error that students make often, too; its high frequency, together with the fact that the system is not always able to spot the error, goes a long way towards explaining the low precision achieved by this preposition. Some examples of errors which go uncorrected are:

(32) I hope, you would provide as usual 10% discount **#RT to [for] /#RT** our company.

(33) We provide internet service **#RT to [for] /#RT** customers.

The verb *provide* is relatively frequent within the data, and it is easy to understand why both humans and computers might be confused by it, as in some cases (e.g. the second example) either preposition is acceptable with no change in meaning.

(34) I can't forget the good things your family don't [did] **#RT to [for] /#RT** me.

In this sentence, the lack of awareness on the part of both student and classifier of the appropriate preposition for the context points to a peculiarity of the English language of which speakers may not normally be conscious: while both prepositions are used for the indirect objects of the verb *do* with a patient role, the choice of preposition depends on whether the action being done is positive or negative. This explains the need for the correction in sentence (34), and perhaps also the classifier's failure to detect the error. Information about the quality of the adjective modifying the noun *things*, which is the only distinguishing element, is not available to it[15]. This finding is useful not just for instructors of L2 English but could also help improve Dapper, if it proved to be a more generalised rule. However, the next example runs counter to these conclusions:

(35) I'm giving this party **#RT to [for] /#RT** Vanya because it's his birthday tomorrow.

The choice of preposition would seem to depend, native speakers suggest, not on whether the occurrence is positive or negative, but on whether the recipient is a direct recipient who has no control over the action (***to***), or is indirectly benefiting from it (***for***). The difference is subtle, and not easy to verbalise. The problems discussed here occur with less frequency in the L1 data. The number of 'misses' on

[15]Compare this to *I can't forget the **bad** things your family did **to** me*, or simply *I can't forget what you did **to/for** me* – a native speaker would automatically understand the implied 'quality' of the things done.

the part of the system is increased by the fact that these examples which DAPPER is unable to correct involve expressions where choosing the appropriate preposition is hard for those who do not have a full grasp of the language, and therefore there is a greater likelihood of such instances occurring in the data.

The issues raised by these examples highlight the difficulties inherent in providing easily generalisable rules and patterns of preposition use, which appear to be highly idiosyncratic not just in relation to the prepositions' heads and objects, but to other items pertinent to those, too. Examples such as these appear hard to solve without recourse to a model incorporating more lexical items, such as the n-gram model proposed by Gamon et al. (2008).

With

With is the preposition with the lowest recall; it is also one of the least frequent in the language, a trend observed both here and in the L1 data. This low frequency makes it, yet again, difficult to understand what may be causing the classifier to overlook errors involving this preposition. Lexical effects due to links between particular items and other prepositions seem to play a role. This means that the low frequency of ***with*** in the language makes it unlikely that there will be many lexical items with a strong association to it: these will be associated with other prepositions instead. Such associations would then override any considerations about the appropriateness of ***with*** in a given context, even when it would be in fact required. An example of this is:

(36) Nowadays, communication is very easy **#RT by /#RT** television, #MD [the] /#MD Internet etc.
Annotators' decision: **WITH**
Classifier decision: **FOR**

Here, the adjective *easy* clearly has a stronger association with ***for*** than ***with*** (cf. *it's easy for me*) which, coupled with the somewhat stilted phrasing, leads DAPPER to give the wrong correction, a pattern observed in several examples. In many other cases, however, the fault does not lie primarily with the preposition's context or the models acquired, but with more complex errors within the sentence, as in:

(37) I've seen last Monday a footboal match from Totenham Hotspurt.
I've seen [saw] last Monday a footboal match **#RT from [with] /#RT**

Totenham Hotspurt **#MV** [**playing**] **/#MV**.
Classifier decision: **AT**

In this sentence, the data available to the classifier, including the lack of the verb *playing*, points it in the direction of ***at***, which is not surprising given sentences such as *I saw him* ***at*** *the lecture*, *I saw a play* ***at*** *the theatre*, and so on. As discussed previously, such sentences should not be considered mistakes on the part of the system as this is the label most consistent with the context presented to it. Arguably, the sentence structure here is rather awkward, so that even the addition of the verb suggested by the annotators, while justifying the choice of ***with***, does not make it an entirely native-like sentence.

Precision also drops greatly. However, as in the case of ***by***, here, too, there is only a small number of instances (just 32) on which to base observations, making it difficult to come to any reliable conclusions. The system's scarce propensity to select ***with*** as a class label was already noted in the analysis of the L1 results. Several of the instances where ***with*** was incorrectly given as the class label are among those which are grammatically correct but show disagreement with the annotators, for example:

(38) I'm very delighted **#RT for /#RT** your proposition.
Annotators' decision: **BY**
Classifier decision: **WITH**

(39) I don't play **#RT at /#RT** the Game boy because I have a computer.
Annotators' decision: **ON**
Classifier decision: **WITH**

These examples involve different prepositions and POS, making it hard to detect any underlying patterns. They also suggest that real precision for ***with*** is higher than it appears. However, this is only half the explanation for the low precision; such a score also means that there are many cases where ***with*** should be given as the class label, but is not – in other words, errors going uncorrected or miscorrected. Again because of the small amount of data clear patterns are not easily spotted. Phrases where Dapper does not recognise an error include several cases involving nouns such as *problem* or *delay*, where the student has written *the problem of [your booking]*, *the delay in/of [your delivery]*, while ***with*** would have been more appropriate instead. It can be surmised that the lexical item is triggering a relation with the incorrect prepositions, preventing Dapper from recognising the error; this can be

easily explained by considering the relative infrequency of collocations with ***with*** as opposed to those with ***of*** or ***in***.

6.3.2.2 Further preliminary conclusions

In the discussion of the L2 error data recall and precision scores, a number of issues affecting the performance of the system were encountered. Generally, there is much overlap with the problems already observed in the L2 correct data task, and to an extent even the L1 data: differences in text type and in the distribution of vocabulary stand out, and certain confusion pairs involving particular lexical items exert a strong influence on the system's decisions. Possible ways of addressing these problems have already been mentioned, such as relying on different types of sources for training, and allowing more than one suggestion to be given.

The system described in Tetreault and Chodorow (2008b) achieves 84% precision and almost 19% recall on preposition error detection, giving an F-score of 31%. The current setup of Dapper, without any further filters applied to the output, achieves 42% precision and 35% recall, giving an F-score of 38%. The system developed by Tetreault and Chodorow includes a set of thresholds defined to strongly favour precision over recall; a similar further tuning of the parameters of Dapper's components is needed to find a better balance between the two measures.

An important observation regards the different distribution of the data here compared to L1 text, which may be having an impact on both recall and precision. Furthermore, the finding that different, non-native-like sentence structure causes problems for the classifier despite being grammatically well-formed is further supported by these results, as many of the system's mistakes can be traced back to stylistic infelicities. This is also something which could be addressed by integrating the training data with other sources such as error-free L2 text, although the model may then become less applicable to L1 tasks.

Human learners and the NLP model were not necessarily expected to run into the same kinds of problems. However, there are certainly aspects of the language which prove problematic for both, as discussed above, for example with regard to locative prepositions and to the use of ***for*** and ***to***. This means that the incorrect L2 data is likely to have a high proportion of instances which the classifier has been shown to be poor at resolving, in either L1 or L2 tasks. Inevitably, this has a negative impact on performance. If further improvements are made to the core model, such problems should not represent as significant an obstacle.

6.4 Determiners

For the determiner task, 2235 instances of NPs were analysed, of which 475 were incorrect in one of three ways: missing determiner ('MD'), unnecessary determiner ('UD') or inappropriate determiner ('RD', i.e. definite instead of indefinite and vice versa). Since one of the three determiner classes is the ***null*** case, DAPPER is able to deal with all three possibilities. Overall accuracy on determiners in L2 writing is **75.53%**, which is an encouragingly high figure; on their dataset, Han et al. (2006) obtain a score of 85%. However, the performance of the system varies greatly between instances which represent erroneous usage and those which do not, so in the discussion that follows the two will be treated separately.

6.4.1 Performance on correct data

Despite the perception that determiner errors are among the most frequent types of errors made by learners, there are usually several nouns within a single sentence, not all of which have determiner errors associated to them. Inevitably, then, the balance of correct to incorrect instances is going to lean heavily in favour of the former. Indeed, this is the case here, and it is immediately clear when the results are analysed more closely that the positive results mentioned above are due almost entirely to the good performance on correct data.

These instances comprise the majority of test instances; on them DAPPER achieves accuracy of **92.85%**, which is slightly higher than the score achieved on correct L1 data. So, unlike the prepositions task, here a change in domain has not affected the system's performance at all, which is a positive outcome. It is likely that this is due to the fact that determiner choice depends more directly on lexical items than on syntactic structure. So, when a text which, as seen, has a different structure and organisation is given to the system for evaluation, such divergences will not affect the determiner component to the same extent.

In comparing L1 and L2 performance, it is also important to assess whether the relative distributions and success rates among the three determiners are maintained; such similarity would be further evidence of the high portability of the system. This information is reported in Table 6.9; the analogous table from Chapter 4, Table 4.15, is reproduced here as Table 6.8 for ease of comparison.

In terms of distribution within the data, we see that the relative frequencies are very similar. There is a slight increase in the frequency of ***a***, accompanied by a decrease for ***null***, but as the sample under examination is not very large this is not

	Proportion of training data	Precision	Recall	F-score
a	9.61%	70.52%	53.50%	60.84%
the	29.19%	85.17%	91.51%	88.23%
null	61.20%	98.63%	98.79%	98.71%
macroaverage		84.77%	81.27%	82.59%

Table 6.8: Individual determiner results - L1 data

	Proportion of data	Precision	Recall	F-score
a	11.65%	78.62%	60.96%	68.67%
the	29.72%	84.91%	92.98%	88.76%
null	58.64%	99.47%	99.05%	99.26%
macroaverage		87.67%	84.33%	85.56%

Table 6.9: Individual determiner results - correct L2 data

a meaningful change. Relative success rates are also unchanged, that is, the classifier is still performing best on ***null*** and least well on ***a***. Indeed, both performance and recall measures for ***null*** and ***the*** display only minimal variation in comparison to their L1 counterparts.

What is more striking is the marked improvement, of around 8%, observed in both measures for ***a***. This would seem to belie the conclusions of the previous chapter regarding the model's less solid grasp on the appropriate contexts for the indefinite determiner, and suggest that it performs fairly well on L2 data, which is after all its intended target. However, the sample size for ***a*** is relatively small (205 instances), so it is not certain that these figures can be relied on entirely; a larger sample may be required to fully support these conclusions.

Are there any peculiarities of learner writing which have a direct influence on DAPPER's performance on ***a***? It is hard to detect any; it is more likely that this is due to a combination of several more subtle features. One of these could be the tendency of students to use phrases such as *a lot* and *a bit* very often, especially when compared to the L1 data used in training. As these are phrases to which the system is able to correctly assign a determiner, this could be one factor which contributes to its better performance. It is also possible that the sentence organisation of L2 writing is simpler and relevant elements of the sentence such as 'there is' phrases (not infrequent in texts which have to describe places or things) are used in a way which more clearly signposts the need for an indefinite article over a definite, simplifying the classifier's job.

Once again, a clearer picture of the classifier's success and its weaker points may be garnered by considering a confusion matrix for the data, which is presented in

Target det	Confused with		
	a	**the**	**null**
a	xx	92.92%	7.08%
the	80.66%	xx	19.34%
null	14.51%	85.49%	xx

Table 6.10: Confusion matrix for L1 data - determiners

Target det	Confused with		
	a	**the**	**null**
a	xx	98.63%	1.37%
the	88.24%	xx	11.76%
null	11.11%	88.89%	xx

Table 6.11: Confusion matrix for L2 correct determiners

Table 6.11; the analogous table for the L1 data, Table 4.18, is also reprinted here, as Table 6.10.

The general trends that emerged in the L1 data are present here, too, in similar proportions. The definite and indefinite determiner are more likely to be confused with each other than with the null case, which in turn is more often confused with ***the*** rather than ***a***. One noticeable difference is the fact that in the learner data there is an even tighter relation between ***the*** and ***a***, that is, the classifier tends to confuse them more often with each other than in the L1 data. This could simply be due to the indefinite determiner's greater frequency in the L2 dataset, or perhaps to a greater use of singular nouns in learner writing: a higher frequency of singular, count nouns would entail a greater likelihood of one of the two determiners being required. In fact, the distribution of singular nouns is almost identical in the two datasets, being higher only by 1% in the L2 data – not sufficient to warrant these differences; equally, the frequency of nouns tagged as 'count' or 'either' is also higher by a couple of percentage points only, so this hypothesis may be unfounded. Overall, though, the finding that DAPPER is able to distinguish between contexts needing a determiner and those which do not with a certain amount of success is of great importance in an L2 context, where most determiner errors are of the 'missing determiner' variety.

In discussing the preposition data, it was noted that not all of the classifier's mistakes are in fact true errors, and that often its choice is acceptable in a different context, or is due to other factors. This is of course the case here, too, and is even more marked than with prepositions since determiner choice is more discourse- and context-dependent than preposition choice. A closer analysis of the incorrect data – which however, given the high success rate of the system, is not a very large

dataset – reveals that indeed just over half (50.86%) of the classifier's choices are not ungrammatical, but simply inappropriate for the particular context, or dependent on other errors in the text.

Three possible explanations for this are considered, which will be discussed in turn: the mistake is triggered by another error; the determiner is grammatical, but inappropriate in that context as the meaning would change; the determiner is grammatical and appropriate in that context, and its choice does not lead to a change in meaning. A fourth possibility, the fact that the system may have identified the wrong element as head noun, was also considered. It was found, however, that this only accounts for a small number of instances, which are likely to depend on errors on the parser's part rather than the model, so it will not be discussed further.

Classifier mistakes related to other errors in the text represent almost 7% of the total. They are mainly of two kinds. The more frequent one involves an error in the number of the noun the determiner is associated with, as in the following examples, where the relevant phrases are highlighted in bold:

(40) I am sure such **a reliable #AGN newspapers [newspaper]** /#AGN #RT of [as] /#RT yours will apologise for this article.
Correct choice: **A**
Classifier decision: **NULL**

(41) It is said that the music played at the Cafe makes **#FN student [students] /#FN** nervous before exams.
Correct choice: **NULL**
Classifier decision: **A**

In both sentences, there is an error in the noun's number, as shown by the correction made by the annotators. The classifier, however, naturally sees the sentence without this correction. In such a context, its incorrect choice is justified and required – cf. ***NULL*** *reliable newspapers* vs. **a reliable newspapers* and *makes* ***a*** *student nervous* vs. **makes student nervous*. But, since Dapper's success is assessed on the basis of its agreement with the text, these instances are counted as errors. However, it is evident that they should not be considered as such, and do not represent a shortcoming on the part of the classifier[16].

In other cases, it is a more complex error on the part of the learner, such as a word order error, that misleads Dapper. An example is given here; for ease of readability

[16]Perhaps a fairer assessment of its success would be on data which does not include any such agreement or number errors.

the incorrect version is given first, followed by the corrected one; again, the relevant phrase is in bold:

(42) Better looks **the picture** at the small shops, the smallest contribution to the total turnover.
The picture looks better for the small shops, which made the smallest contribution to the total turnover.
Correct choice: **THE**
Classifier decision: **A**

The issue here revolves around the NP at the start of the sentence, *the picture*. Although it is impossible to know precisely why the classifier made its incorrect choice, the unusual and non-native like structure of the sentence initial clause may well play an important role. For example, it could be that this particular sequence of POS tags is more compatible with an indefinite than a definite determiner, especially near the start of a sentence. Again, arguably such mistakes should not be counted against the classifier, as they occur within phrases which are involved in other errors.

Next, instances where the classifier's choice is both grammatically and pragmatically correct are addressed. In the discussion of the prepositions task, a significant number of such cases was found; here, they amount to around 11%. This is in part suprising, as one might not expect there to be so many cases where either definite or indefinite article are equally appropriate, given that usually the two determiners do give a different meaning or connotation to the text. It is therefore interesting to consider some examples; the relevant phrases are in bold.

(43) I was warned by **an agency** to be pecient [patient] and careful while doing this job because the owner of this house was #MD [a] /#MD peculiar old man.
Correct choice: **A**
Classifier decision: **THE**

(44) **In winter** people are often to use #UD the /#UD electricity in the night time.
Correct choice: **NULL**
Classifier decision: **THE**

(45) **A second point** against the theory of genetics is that it depends on #R their theory on the /#R #W twins behaviour [behaviour of twins] /#W.

Correct choice: **A**
Classifier decision: **THE**

In these sentences, the determiner suggested by DAPPER, while marked as an incorrect choice because it is not the same as that used in the original sentence, is grammatically correct and its use instead of the original does not appear to bring about a change in the content of the statement. Indeed, in the first sentence especially, it could be argued that the use of a definite article is more appropriate and native-sounding. In the second and third sentences, too, the presence or absence of the determiner brings no difference in meaning. These examples, and the others like them, should therefore not be considered mistakes on the part of the classifier, especially where, as shown, the classifier's suggestion may be more natural than the annotators' judgements.

We come finally to the largest subset of this group of instances (30%), namely those where the classifier's choice, although grammatical, would cause a change in meaning and is therefore not acceptable. This is almost inevitable in the domain of determiners, and indeed is an issue encountered in working with L1 data: there, in carrying out an error analysis, it was observed (Section 4.3.2) that it was often impossible to understand why the classifier choice was inappropriate without being able to contextualise the sentence within its wider discourse. Indeed, the examples that follow will help illustrate this point: we can agree with the original text in finding the classifier choice wrong because of our knowledge of the world, but there are some cases where the difference is subtle and it is no surprise that both human learners and DAPPER find them confusing. Again, some errors are not annotated, and relevant phrases are in bold.

(46) A few weeks before, you underlined in **a newspaper article** the fact that sport is not so popular at this moment.
Correct choice: **A**
Classifier decision: **THE**

(47) In order to help us to popularize sport amongst the youth, I #TV will be [am] /#TV pleased to invite you to **a competition** organized by the International Students' sports club.
Correct choice: **A**
Classifier decision: **THE**

In these two sentences, either determiner would be correct grammatically, but pragmatically, the choice depends on whether it is the first time the relevant noun (*news-*

paper article, *competition*) is being mentioned. Without knowing the wider context, one can nevertheless assume that the annotators did not mark these instances as incorrect because they are indeed the first mention of each; this is the kind of classifier error that is hard to avoid without access to the rest of the text. There are several such instances in the data.

(48) The reception area of the hotel was strangely silent and even though he rang **the bell** several times, nobody answered.
Correct choice: **THE**
Classifier decision: **A**

Here, the classifier's choice of indefinite determiner could be acceptable and would not change the content expressed by the sentence. However, the use of the definite determiner sounds more native-like and appropriate, due to our knowledge of what hotel receptions are like: a bell is typically found there, and so can be considered a unique referent, for which the definite determiner is required. This type of classifier mistake, of which there are quite a few, is caused by lack of world knowledge on the part of the system, which would be hard to include in the model.

(49) Luckily enough for Mary, he had always wanted to be **a father**.
Correct choice: **A**
Classifier decision: **THE**

This sentence is similar to the previous one in that, despite both determiners being grammatical, the correct choice depends on conventions of the English language rather than previous mentions within the discourse. It is a general tendency of English that the indefinite article must be used when referring to a role or profession in the abstract (cf. for example *I want to be **a** teacher when I grow up, her father is **a** doctor*). This is a linguistic fact rather than something dependent on external knowledge, and as such it is the sort of regularity that one would have expected to be picked up by the model, but this seems not to be the case. In conclusion, various causes for these mistakes which lead to grammatical sentences with different meanings have been identified; at the current state of development, overcoming them is challenging, but some possibilities have been suggested in the course of this chapter and in Section 4.3.2.

The discussion above of 'easily explained' classifier mistakes has revealed that not all of them should be counted as real mistakes. If one were to exclude from the count the mistakes due to other errors and those where the chosen determiner does not lead

to ungrammaticality or a change in meaning, the accuracy figure on correct instances would rise to **94.15%**, which is a positive result.

As for the instances which are clearly errors on the part of the classifier, and cannot be ascribed to other factors, it is hard to say what is misleading DAPPER into making an incorrect class assignment. Some examples are shown here, but it is clear that there is no easily identifiable pattern to the erorrs.

(50) As an engineer he would be **the perfect person**, because he has the needed [necessary] knowledge about steel.
Correct choice: **THE**
Classifier decision: **A**

(51) A few years ago I was offered a part-time job as a housekeeper in **a beautiful house** in the north of London.
Correct choice: **A**
Classifier decision: **THE**

(52) Not **a soul** appeared.
Correct choice: **A**
Classifier decision: **THE**

(53) **The special knowledge** about our steels and firms will be learnt on different courses.
Correct choice: **THE**
Classifier decision: **A**

These are just four examples of instances where the classifier's choice leads to an awkward-sounding or outright ungrammatical sentence. It must be concluded that inappropriate links between contexts and class labels have been created by the system: this is evident for example in the last sentence, where the noun *knowledge* is identified as a mass noun, but is nevertheless paired to an indefinite determiner, which is not what one would expect. In the third sentence, the grammatical construction *not a X* is very distinctive, so it is surprising that the classifier is not able to recognise it. It can be hypothesised that the effect of a general NP such as *not **the** X* (*it is **not the** right answer, it is **not the** car we need*, and so on) overrides this distinctive pattern by virtue of being more frequent.

In concluding this section, it must observed that although it has not always been possible to offer an explanation for all of DAPPER's mistakes, it is important to

remember that these constitute only a small percentage of all instances submitted to the system. Overall performance on L2 correct instances is comparable to that on L1 data, and unlike in the case of prepositions, the characteristics of L2 writing do not seem to have a strong negative effect on the results. Precision is between 78.6% and 99% for the three determiners, and recall between 61% and 99%, suggesting this is a tool which can be used with a high degree of confidence in this context.

6.4.2 Performance on incorrect data

We now turn to an analysis of the results on incorrect data, i.e. cases where DAPPER is required to recognise and correct errors involving determiners. As anticipated above, performance on erroneous data is much lower than on the correct data: overall accuracy is just **7.93%**, although there are interesting differences among the recognition rates for the three types of errors. Again, exact figures and distributions cannot be given for intellectual property reasons, but missing determiners tend to account for the majority of error instances, followed by unnecessary determiners, and then wrong ones[17].

6.4.2.1 Wrong determiner

The discussion begins with the 'wrong determiner' ('RD') types of errors, that is, use of ***the*** when ***a*** is correct and vice versa. These are the least frequent but also the ones which are dealt with successfully most often: accuracy on these is 66.7%. This is an encouraging result; typical errors which are corrected by the classifier are given here:

(54) Sadly, while he was at the top of his form, **#RD the /#RD sudden illness** struck.
Annotators' decision: **A**
Classifier decision: **A**

(55) Rent your accommodation near the university or near the public transport, but not in **#RD a /#RD center** of the city, because of huge expenses.
Annotators' decision: **THE**
Classifier decision: **THE**

[17]It should be noted that in the corpus, determiner errors also include those involving possessive pronouns such as *his* or *their* and demonstratives such as *this* or *that*, which are not included in this study; it is possible that within those classes distributions differ.

(56) Regarding the opening hours, **#RD the /#RD high percentage** of the customers think they are good.
Annotators' decision: **A**
Classifier decision: **A**

Although none of these determiner errors impair meaning in a serious way, they nonetheless sound wrong or awkward. The classifier's input renders the sentences grammatically correct and more native-like, and could be the starting point for a reflection on the use of definite vs. indefinite determiners. As for the errors that DAPPER misses, these are due to a similar variety of reasons as discussed above in relation to the correct instances. Over half are simply due to pragmatic or stylistic issues, two examples of which are given here:

(57) So I decided to join **#RD the /#RD** English Class.
Annotators' decision: **A**
Classifier decision: **THE**

(58) I saw it in **#RD the /#RD** magazine.
Annotators' decision: **A**
Classifier decision: **THE**

As already discussed in the previous section, in cases like these it is very hard to know what determiner is required without knowledge of the wider context. In a small number of cases – less than 8% – the classifier suggestion, which coincides with the original determiner used by the learners, is to be considered equally acceptable, cf.:

(59) The turnover for heaters has been continuously decreasing, from 16% in 2000, to 14% in 2001 and up to **#RD a /#RD** spectacular decrease of another 7% in 2002.
Annotators' decision: **THE**
Classifier decision: **A**

Here, the classifier suggestion is ***a***, so it does not register the presence of an error; the text does not seem to change in meaning or grammaticality by using the indefinite determiner. Other missed errors (15%) are due to the presence of other errors, for example:

(60) So if you want to speak English like **#RD the [a] /#RD #AGN natives [native] /#AGN** #UV do /#UV, be sure to take part in this programme.

Annotators' decision: **A**
Classifier decision: **THE**

This is analogous to examples already observed in the previous section. The annotators have marked the determiner as incorrect as a consequence of there also being an error in the number of the noun. Without their correction of the noun, the choice of ***the*** as determiner in this context – made by both DAPPER and the student – is perfectly acceptable, indicating that these types of mistakes on the classifier's part should not be included among its flaws.

The remaining 20% or so of mistakes, on the other hand, are true mistakes, the possible causes of which, as already discussed in the context of the correct instances, are hard to pinpoint. It is interesting to see, however, that while in some cases the error is simply not detected, in others – such as the first example below – the classifier is partially successful, as it does recognise the inappropriateness of the determiner used, but then suggests another incorrect alternative in its place. It could be argued that having a system which alerts one to the presence of an error can still be of great value, as it could lead the user to consider their choice more carefully. On the other hand, if the error recognition stems from the classifier's assumption that another incorrect determiner is required, the potential for misleading learners remains high.

(61) It is not **#RD the /#RD pentium**, but it is an IBM 486.
Annotators' decision: **A**
Classifier decision: **NULL**

(62) Worst of all, it may cause **#RD the /#RD conflict** between the races.
Annotators' decision: **A**
Classifier decision: **THE**

By excluding the instances where DAPPER's mistakes are due to other factors, or result in an equivalent sentence, accuracy on this task rises to 75%.

6.4.2.2 Unnecessary determiner

In terms of frequency, 'unnecessary determiner' ('UD') is the next most frequent type of determiner error found in the data. Unfortunately, it is also a type of error which DAPPER finds difficult to correct, doing so successfully in just over 3% of cases. An example of a well-corrected instance is:

(63) Phone me #RT at [on] /#RT **#UD the /#UD** 06.92.33.07.
Annotators' decision: **NULL**
Classifier decision: **NULL**

But what is making it so hard for almost all errors to be detected? A small proportion – just over 9% – belongs to the group of instances where the classifier's decision does not cause a crucial change in meaning:

(64) I am a bit shy, and it is difficult for me to mix with **#UD the /#UD others**.
Annotators' decision: **NULL**
Classifier decision: **THE**

(65) During **#UD a /#UD summer time** we've got a lot #MT [of] /#MT attractions in the centre.
Annotators' decision: **NULL**
Classifier decision: **THE**

Sentence (65) is interesting not just because it is a good example of a case where the classifier's suggestion is arguably better than the annotator's correction, but also because it further highlights an issue already introduced in the previous section: cases where Dapper correctly identifies the inappropriateness of the determiner used by the student, but then suggests an erroneous alternative. In this task, such 'half-way corrections' are very frequent – around two thirds of cases. They are considered full mistakes on the part of the classifier, as arguably it is not fulfilling a useful function if it is giving learners misleading and incorrect advice. On the other hand, its ability to recognise the incorrectness of the determiner present is an important part of the error recognition process, and should not be discounted lightly: it suggests that if the mechanisms for the suggestion of alternative choices to the student were improved, this component would prove much more successful.

These results do, however, cast doubt over previous claims about the system's strong model for the ***null*** case. Surely, if the model for this class is so reliable, it should not be difficult for the classifier to know when it is required, and to therefore identify instances of unnecessary determiner errors? It can only be hypothesised that the contexts of these errors are not as straightforward as expected, which could also account for difficulties encountered by the learners.

Analysing these classifier mistakes further, it is found that around 12% are directly linked with other errors in the text, which, as argued above, makes Dapper less

directly responsible for them. Many of these are of the type already encountered in other contexts, namely an error made by the learner in choosing the number of the noun, whose correction renders the original choice of determiner incorrect:

(66) I know you really love reading **#UD a /#UD #FN book [books] /#FN**.
Annotators' decision: **NULL**
Classifier decision: **THE** (error detected but wrong correction given)

(67) Hence, **#UD a /#UD good interpersonal #FN skill [skills] /#FN** #AGV is [are] /#AGV also a child's achievement.
Annotators' decision: **NULL**
Classifier decision: **A** (no error detected)

In other cases, however, there is a more complex underlying error, as in these examples:

(68) Also when the television is on it might interrupt you many other times and you will lose **#UD the /#UD interes [interest]** in the book.
Annotators' decision: **NULL**
Classifier decision: **THE** (no error detected)

The classifier's error in this case is likely to stem from the misspelling of the noun *interest*. The parser interprets this word as the plural form of a non-existent **intere*, a term for which of course the model has no information. It must therefore rely on other contextual cues to make its choice: perhaps with the correct spelling of the word, it would have been more successful. Such problems can be avoided if, as already suggested, a spellcheck filter is used on the data before error detection.

(69) I would like to know if you need **#UD a /#UD #FV book [to book] /#FV** accommodation.
Annotators' decision: **NULL**
Classifier decision: **A** (no error detected)

In this sentence, the error in the verb form (the missing *to*) misleads the parser into interpreting *book accommodation* as a NP (cf. *the holiday accommodation, a book expert*). So, even though *accommodation* normally would not occur with an indefinite determiner, in this particular context it would not seem so awkward. This is the kind of mistake which is not unlikely to occur when working with L2 text, and it is hard

to overcome as it is not always possible to predict what kinds of errors a particular stretch of writing will contain.

A large proportion of classifier mistakes – over 23% – falls under the 'context-dependent' category, which, as discussed in other cases, is hard for the classifier to perform well on without access to discourse information. Some examples are given here to illustrate the types of contexts this problem is encountered in:

(70) Buy **#UD a /#UD** cheap food products.
Annotators' decision: **NULL**
Classifier decision: **THE**

(71) However, I think parents need to tell their children how to spend **#UD the /#UD** money.
Annotators' decision: **NULL**
Classifier decision: **THE**

In these examples, as in other similar instances, the annotators know that the determiners highlighted are unnecessary and incorrect because the student is making a general statement about food products or amounts of money as abstract entities rather than referring to specific instances of them, but this is not something DAPPER, or indeed a reader presented with the individual sentences in isolation, is able to discern.

Finally, over half the classifier's mistakes (54.6%) are simply true mistakes and cannot be explained easily, as the examples below show:

(72) Capitalism, on the other hand, has changed a lot throughout **#UD the /#UD history**.
Annotators' decision: **NULL**
Classifier decision: **THE**

(73) I #RV stayed [spent] /#RV all **#UD the /#UD day** #RT at [on] /#RT the sandy beaches.
Annotators' decision: **NULL**
Classifier decision: **THE**

(74) Therefore, **#UD the /#UD cultural communication** will benefit a multi-cultural world through mutual understanding.
Annotators' decision: **NULL**
Classifier decision: **THE**

Clearly in these cases there are patterns of usage where the phrases and collocations involved have been acquired imperfectly or not at all. In the first sentence, for example, *throughout history* is not a very unusual phrase, and it is surprising that the error in it is not spotted. Similar considerations apply to the second example, perhaps even more surprising as *all day* is also a simple phrase, and a very distinctive one. It can be hypothesised that perhaps here a text type effect is at work, such that these simple expressions typical of narratives do not occur as frequently in the BNC as they do in learner essays. This explanation however is less plausible for the third example, since *cultural communication*, or at least just *communication*, can be thought of as a phrase which is not so uncommon in the types of text found in the BNC. The conclusion that must be drawn is that there are simply some contexts where the system does not yet perform adequately.

To sum up this analysis of DAPPER's performance on recognising and correcting errors involving an unnecessary determiner, it is claimed that by excluding those instances where its mistakes are due to external factors, or result in a sentence of equivalent meaning, accuracy on this task rises to almost 23%. While this is still not a good result – mere random guessing would have yielded a higher score – it is a marked improvement on the first score reported. If some of the issues discussed here are addressed, this component will become of increased usefulness.

6.4.2.3 Missing determiner

Finally we come to the most frequent kind of determiner error, namely missing determiner ('MD') – instances where the learner is not aware that one of ***the*** or ***a*** is required. Since the model's performance on these two classes is not as high as it is for the ***null*** class, results on this task might be expected to be not very high. Indeed, accuracy is the lowest of the three tasks: just **1.5%**. An example of a typical error, in this case well corrected by DAPPER, is given here:

(75) To attract more customers #**MD [a] /#MD wide range** of products have to be #RV brought [bought] /#RV.
Annotators' decision: **A**
Classifier decision: **A**

The classifier's poor performance on this task is particularly disappointing because of the high frequency of this kind of error in learner writing; it is therefore important to analyse the mistakes made to understand where the system can be improved. Unlike the previous two tasks discussed, here 'explainable mistakes' – those where

the classifier's mistaken choice is not necessarily to be interpreted as such – account for only 16.5% of the total number of mistakes. These cases will be considered first, although it is clear that the greater share of the problem lies where it cannot be easily addressed. We begin with the 4% of instances where the suggested correction is also acceptable without bringing a change in meaning, as in these examples:

(76) There is a clear comparison between the four selected countries (Australia, Japan, **#MD [the] /#MD United Kingdom** and **#MD [the] /#MD United States of America**) in three aspects of the water consumed.
Annotators' decision: **THE** (in both cases)
Classifier decision: **NULL** (in both cases)

(77) On the other hand, I and many #UT of /#UT people believe that **#MD [a] /#MD personality** is not born with us.
Annotators' decision: **A**
Classifier decision: **NULL**

In these examples, DAPPER agrees with the learner's original choice of not using a determiner, which the annotators believe to be wrong. As observed before, there are a number of contexts where more than one determiner may be appropriate without a change in meaning, and the examples above are two such cases. In the first sentence, the annotators' correction may seem surprising given that the country names are in a list rather than occurring as subject or object of a phrase, where indeed a determiner would have been necessary[18]. As for the second example, leaving aside the issue of its slightly stilted prose, arguably the sentence sounds better without the determiner, as it is meant to be a general statement[19]. So once again it is found that not all cases where DAPPER disagrees with the annotators are actual mistakes on its part.

As for instances where the classifier choice is grammatical, but causes a change in meaning, these account for around 6% of mistakes. The examples given below illustrate once more how difficult it is to know what determiner is appropriate when more is not known about the context. In all cases, the classifier's choice was simply ***null*** - that is, it did not detect the presence of any errors.

(78) I would strongly advise you, **#MD [the] /#MD organisers**, to find a solution concerning the use of computers.

[18] Although not all native speakers agree that this sentence would be acceptable without determiners.

[19] At most, a possessive such as ***our*** *personality* would have been a more appropriate correction.

(79) The guard on the train came and checked **#MD [the] /#MD tickets**.

(80) They have **#MD [the] /#MD skills** and experience to develop the products.

These sentences illustrate different ways in which context impinges on determiner choice. In the first case, it would indeed be more common to find a determiner in that phrase. However, it is not impossible to envisage a wish for a rhetorical flourish whereby the determiner is not needed, and a more direct appeal is made to the *organisers*. The second sentence falls under the category of 'lack of world knowledge'. While it is possible to say just *checked tickets*, it is understood, and part of what we know about the world, that one generally requires a ticket to travel on a train, so that *the tickets* in this context are a unique referent and a definite article is more appropriate even if they are being introduced for the first time to the discourse. As noted before, this kind of mistake is hard to avoid because of its dependence on external, non-linguistic factors. Finally, the third sentence is a simple case of a choice that is inappropriate in this context because here a particular set of skills and experience is being picked out. Proficient speakers of the language can infer this from the phrase that follows, which restricts the range of possible referents, making a definite determiner necessary, but it is not clear if this information is acquired and used by the model.

Finally we turn to mistakes due to other errors in the text; these account for almost 7% of the total. Several are of the kind already encountered, namely where the determiner error follows on from the identification of an error in the number of the noun; cf.

(81) But #RA they [it] /#RA #AGV were [was] /#AGV **#MD [a] /#MD great #AGN holidays [holiday] /#AGN**.
Annotators' decision: **A**
Classifier decision: **NULL**

The sentence, which is grammatical in its plural form, has been corrected by the annotators so as to be in the singular rather than in the plural. This change leads to the need for a determiner which was previously unnecessary, but DAPPER is not to be blamed for not perceiving the need for an ***a*** when the relevant head noun is in the plural – in fact, it would be a serious shortcoming if it did.

(82) #R After the end of cards production /#R, **#DV delivery [deliver] /#DV #MD [the] /#MD cards** by car to my office.
Annotators' decision: **THE**
Classifier decision: **NULL**

Sentence (82) introduces a different issue, namely an error on the part of the learner where the wrong POS has been used for a lexical item, which has the effect of leading the system to misidentify the boundaries of the NP. The use of *delivery* rather than the required verb *deliver* means that the sentence is read as containing the NP *delivery cards*; despite this being nonsensical in the context, the classifier defaults to its choice of ***null***, presumably precisely because it cannot make sense of the sentence structure.

Learner errors are not just restricted to misused POS and agreement, however. Missing words can also be an issue, as in the example below:

(83) I have decided to reward all the staff by 5%.
I have decided to reward all the staff #RT by [with] /#RT **#MD [a] /#MD 5% #MN [pay rise] /#MN**.
Annotators' decision: **A**
Classifier decision: **NULL**

With the addition of the noun *pay rise*, and the change of the preposition from *by* to *with*, the sentence requires the use of a determiner. However, these are all changes that the classifier is not aware of, because the text is stripped of all error annotation before it is submitted to the system. In the original version of the sentence, the absence of the determiner does not come across as erroneous, so it is not surprising that Dapper did not register the presence of an error.

The analysis above has given further examples of the types of problems encountered in identifying and correcting determiner errors. However, as noted, the majority of missing determiner errors simply go unnoticed without any clear triggers for these mistakes emerging. Some examples of uncorrected errors follow, to show the variety displayed. In all cases, the classifier's decision is simply ***null***; in other words, the instances are incorrectly found to be free of error.

(84) **#MD Few [A few] /#MD weeks** ago, you underlined in a newspaper article the fact that sport is not so popular at this moment.

(85) There #AGV are [is] /#AGV **#MD [a] /#MD range** of snack food as well.

(86) The general awareness #M [of appearance] /#M among the masses has also increased many times in **#MD [the] /#MD recent past**.

(87) They were based on **#MD [the] /#MD trading market** and gave us more experience with #RV dealing /#RV.

As already observed, many of these classifier mistakes occur in relatively simple, frequent phrases. Their frequency would suggest that statistical regularities for them should be easy to acquire; on the other hand, it could also have a negative impact, in that they could occur in so many combinations as to make generalisations impossible. For example, it would be highly unlikely to find temporal expressions such as *in the past* (cf. (86)) without a determiner. In fact, it emerges that this particular phrase is not so frequent in the BNC data, occurring only 260 times. This points to the fact that the difficulties raised by differences in text-type discussed throughout this chapter may also affect words and phrases which might not at first sight appear problematic.

Phrases involving *few* are more challenging as we can say both *a few X* and *few X*, although this brings about a shift in meaning (cf. *few people came* vs. *a few people came*). In a temporal phrase such as the one in the first sentence, however, the determiner-less option would not be available or grammatical. Mistakes such as the one in the second sentence are also unexpected since the system correctly identifies *range* as a singular count noun, so the absence of a determiner associated with it should result in an error[20]. An overview of the data reveals that instances of both missing determiners are unrecognised with equal frequency.

So far differences between learner texts and the BNC have not been found to be a major factor in impairing performance, but there is one aspect in which the two differ which comes to the fore in analysing the results of this task. It is not so much a difference in topics as in age of the texts: the BNC was created with material from almost 20 years ago, and learners writing in contemporary English may use lexical items which are not frequently seen in the BNC. In this task, the lexical items play a crucial role in class assignation. If the noun in question has not been seen in training, the classifier may be unable to make an informed choice as it cannot rely on any information associated with it.

A clear example of this discrepancy is the noun *internet* (cf. also the discussion in Section 6.3.1), which requires the definite article in English, but not in several other languages, leading to countless sentences such as *I saw it in internet*, *I booked it on*

[20] And indeed it is dealt with correctly in (76).

internet, and so on. This is one of the 'missing determiner' errors in the CLC which the model never detects: a fact which is not surprising when one notes that this noun occurs only four times in the whole of the training data. It has been observed in Section 6.3.1.1 that using different sources of training data may help overcome some of the issues found in the classifier's performance; this is a further example of a problem that could also be addressed in this way.

In light of this discussion, a revised figure for DAPPER's accuracy on the missing determiner task can be obtained. Excluding those cases where its mistakes are due to other elements of the sentence, or result in a sentence of the same meaning, accuracy rises to almost 13%. This is still a clearly unsatisfactory result, and more work is needed to ensure the usefulness of this component.

6.4.3 Further considerations

As observed in the course of this chapter (see especially Section 6.3.2.2), the fact that DAPPER is still in the early stages of development, together with the differences in datasets used, makes it difficult to assess its performance directly against that of other systems, such as Han et al. (2006) or Gamon et al. (2008), already introduced in Section 2.3.1. The former paper reports results on a very large body of text – over 20,000 NPs, of which 2700 contain errors. In their data, the distribution of error types is found to mirror that of the present work, highlighting the lower frequency of ***a/the*** confusion errors in comparison to missing or unnecessary determiner errors. The first reported accuracy score of 85% is very high because of the large number of correct instances present in the data; a clearer evaluation of that system's performance can be had by focusing on the error component only. Several different techniques are used by the authors to improve error detection, and in many cases, missing determiner errors are not accompanied by an indication of which determiner is required. Overall, precision on missing and unnecessary errors is between 80% and 90%, with recall lower at around 20-40% depending on the particular configuration of the system. On ***a/the*** confusions, the figures are much lower, with precision at 67% and recall at 11%. The greater difficulty of detecting confusion errors is in contrast to what is observed in DAPPER's performance, where these are detected with greater ease; it is possible that the contextual predicates chosen somehow favour the acquisition of these types of contexts over others. Overall precision achieved by DAPPER on error detection (all three types) is only 22%, so arguably its determiner correction component is not as competitive as its preposition counterpart.

The system described in Gamon et al. (2008) achieves 59% accuracy on determiner error detection and correction. No details are given about differences in performance among the three error types, so a more detailed analysis cannot be carried out. Missing determiners are found to be by far the most frequent type of error encountered. The comparatively high accuracy score achieved by the authors suggests that their system must be detecting a significant number of these, in contrast to Dapper, for which missing determiners are hard to identify.

As a general consideration regarding the distribution of determiner errors, informal observation of the data shows that all three kinds of errors occur both in otherwise correct sentences and together with several other errors. This confirms the impression that determiner use is one of the hardest areas for learners to master, even when their command of the grammar is fairly advanced.

6.5 Final observations

This chapter set out to answer two questions: what is the general effect of porting NLP tools to the domain of L2 writing, and is it viable to use them in an error correction application? The first issue has been addressed by examining the performance of Dapper on L2 writing which did not contain any preposition or determiner errors. On the preposition task, accuracy is around 10% lower than on its L1 counterpart, while no such differences are found in the determiner task. These results suggest that there is great potential for the use of NLP tools, even those consisting of several components, on learner writing. The detailed comparison of individual results to those obtained in the L1 tasks yielded several insights into differences between the two text types which do raise problems, such as:

- Concomitant errors: presence of more than one error can mask intended meaning
- Content differences: differences in topic or age of texts
- Stylistic differences: style and sentence structure of L2 writers are not always L1-like

The identification of these issues, and of possible ways to address them, as discussed in the course of the chapter, will play an important role in the future development and improvement of Dapper.

The application of DAPPER to texts containing preposition and determiner errors aimed to establish the viability of using a classifier-based model trained on L1 text to correct errors in L2 text. On this task, the success of the system is limited. Around 40% of preposition errors are corrected appropriately, with an average precision of almost 42% and average recall of 35%. Determiner errors are currently not dealt with sucessfully by DAPPER: while confusion between ***a*** and ***the*** is corrected in around 67% of cases, missing and unnecessary determiner errors are hardly ever detected. More research is needed to establish whether this is due to the inherent difficulty of the task, or to shortcomings of the model which make it ill-suited to solving this problem.

Certainly, the difficulties typical of a generic error detection task are compounded by challenges posed by the characteristics of L2 text. Currently, DAPPER evaluates text to which no changes or corrections have been made, including all spelling errors, which no doubt is a factor in its less-than-satisfactory performance. The possibility of introducing measures to facilitate the tools' work on the data, if these appear to be necessary, has been envisaged. For example, spelling or some non-preposition or determiner errors can be corrected to see which of the other errors are worth focusing on. However, removing all but the preposition and determiner errors is not advisable: this would create an artificial type of L2 text which is hardly ever encountered, and would therefore not constitute a representative setting for this task.

In conclusion, the results and analyses presented in this chapter have shown that a model-based approach to error identification is viable, but further work is needed to ensure it is more attuned to the specific requirements of L2 writing.

Chapter 7

Conclusions and future directions

The contributions of the research presented in this thesis are briefly summarised, together with proposed avenues of future work.

7.1 Conclusion: thesis contributions

In Section 1.4, the main contributions of the present work were outlined. These are referred to again in the course of this section.

In Chapter 3, the development of Dapper was outlined, motivating the choice of features selected as representative of prepositions' and determiners' contexts; and explaining the procedure used to create the feature vectors representing these contexts. The model was tested on L1 data, and the results of these tasks were presented in Chapter 4. Accuracy scores achieved on selection tasks for prepositions (70.06%) and determiners (92.15%) are shown to be not dissimilar to other related work in the field. Furthermore, the scores indicate that this type of classifier-based approach can successfully be used to acquire models of use for these two POS.

Chapter 4 also presents an in-depth analysis of results on each preposition and determiner, with reference to the ease with which models of use for individual items can be acquired, and to particular relations which may be detected between specific items and features. There are very few studies of this kind, and the treatment of this material as the basis for resources of use to the research community is envisaged, as described in the next section.

The role played by each feature was the topic of Chapter 5, where the contribution of each feature to overall performance was analysed to quantify its statistical and linguistic import. Although most researchers who approach these tasks devote considerable attention to the issue of feature selection, this thesis gives scope for a fuller analysis of the topic. This allows for a number of considerations which may be

of interest to research in the fields of NLP, linguistics, and L2 English teaching. The analysis is particularly important as one of the aims of this work is to assess the role of more sophisticated syntactic and semantic features within the model. Lexical items and POS sequences are found to be among the most important factors in preposition and determiner occurrence; however, the use of syntactic analysis remains central, to ensure that the correct items are identified as belonging to the relevant PPs and NPs.

Finally, the performance of Dapper on L2 text was evaluated in Chapter 6. It was found that on texts where there are no preposition or determiner errors, performance does not degrade by more than 1% in comparison to L1 data, suggesting the models developed are sufficiently robust to be applied to a different domain. Results on error correction are promising, but still require refinement. Comparisons between results on individual items in L1 and L2 tasks are also made. These yielded several observations on the application of NLP tools in the L2 domain which can enrich the ongoing discussions in the field regarding the best ways to make use of them. The trends that emerged from these studies can also be of assistance to L2 instructors, drawing their attention to areas of particular difficulty or confusion for learners.

7.2 Future work

Throughout the course of this research, two factors have emerged as requiring particular attention if Dapper's performance is to be improved: differences in text type between training and testing data, and the presence of contexts where more than one choice may be appropriate. The issue of domain differences between L1 and L2 data has been discussed at some length in Chapter 6, where the possibility of using other sources of data in training was suggested. While this may not make much difference to results on L1 tasks, it might prove beneficial in the context of L2 error detection. The next step is therefore to train models which include data from corpora whose content is more comparable to that of the learner essays. An obvious candidate is LOCNESS, the Louvain Corpus of Native English Essays[1], which contains argumentative and expository essays, similar to the kind found in learner corpora. If results improve significantly with this modification, the validity of the approach at the heart of Dapper will be confirmed: it can then be considered a general-purpose tool that can be adapted to a variety of domains.

Deciding how to choose among more than one plausible preposition or determiner option is more challenging. At present, Dapper's setup allows the choice of one

[1] http://cecl.fltr.ucl.ac.be/Cecl-Projects/Icle/LOCNESS1.htm

answer only, regardless of whether there may be others with similar likelihood. It is possible, however, to enable the output of more than one choice, for example the two or three that rank as most likely. Early investigations into the feasibility of this enhancement would seem promising: on the L1 preposition task, accuracy increases by over 12% to 82.75%. On L2 preposition error instances, 57% of errors are detected, compared to the previous figure of 39%. In both cases, 2-3 labels are output on average. While careful tuning of parameters will be required to ensure that such higher rates of recall are not coming at the expense of precision, the figures suggest that this kind of approach must be considered if DAPPER is to perform more reliably.

A problem arises from having a multi-item output, however: knowing which item to select. The choices are ranked, and a confident user could rely on these rankings, and their own knowledge of the context, to infer which item is most likely to be appropriate, but this might prove too challenging for less proficient speakers of the language. One solution could be to integrate the system with an additional component which draws its knowledge from the world wide web. After all, it is common for people to use web data as an informal source of information and clarification of doubts (for example when unsure between two forms, querying a search engine for both and choosing the one with the greater number of results). Web data would be better suited to the preposition task than to determiners, since the latter are so context dependent that, apart from clear-cut cases involving mass nouns, it would be almost impossible to find reliable answers for one case or the other.

One way of harnessing the benefits of web data is to use the the Google n-gram Corpus introduced in Section 3.1.4.1. The n-grams surrounding the target preposition or determiners can be identified and submitted to a component which issues a query to the corpus for combinations of the phrase with each of the possible prepositions or determiners. The frequencies of each n-gram sequence are returned, and this information can be used by the learner to gain a clearer understanding of the usage of each item. Additionally, it can be used to integrate the classifier results when a lack of strong contextual patterns prevents it from making an informed choice. A very small preliminary study investigating the implementation of this module is reported in De Felice and Pulman (2007), where it was found that the two approaches – feature-based and web data look-up – complemented each other well. More recently, both Gamon et al. (2008) and Tetreault and Chodorow (2008b) have experimented with similar additions to their model. The latter also rely on the Google n-gram corpus, but report that its introduction brings only a negligible improvement to results. The former paper also envisages using web results as a source of examples for the user,

but the contribution of this component to their model has not yet been assessed. The benefits of the addition of web data are therefore not clear. In contrast to these findings, Lapata and Keller (2005) use web counts for a variety of tasks including determiner generation and conclude that this approach outperforms the use of counts from the BNC. It may be that the feature-based models developed, including DAPPER, are sufficiently informative so as not to require further sources of information; alternatively, the best implementation of this data source may not yet have been identified. In either case, the issue warrants further investigation.

Finally, the Introduction suggested the possibility of making further use of the data on contextual patterns, feature frequency, and preposition and determiner occurrence that has been acquired. This could take many forms, ranging from simple guidelines and advice to students (in the form of a website or text) to a more sophisticated resource such as a database in which information such as lexical and syntactic preferences can be stored – both for prepositions/determiners and for lexical items. A hypothetical entry for the verb *drive* could state, among other things:

> DRIVE, verb. Typical occurrence frames: drive to, drive in, drive by.
> ***To*** is used when the object of the preposition is the location of your destination, often a proper noun; example: *John drove to France for a holiday.*
> ***In*** is used when the object of the preposition is the item you are driving in, or the location where the driving action is taking place; example: *Driving in Italy scares foreigners.*
> ***By*** is used when the verb occurs in the passive form. The object of the preposition is often an abstract entity; example: *She was driven to madness by her fear of goats.*

The example above is quite verbose, and modeled on those found in learner texts, as described for example in Chapter 2, but of course a machine readable form can also be easily envisaged. Although similar resources do exist which cover some or most of the aspects discussed above, the extensive coverage afforded by the data collected for the present research would make this a useful contribution to the community.

In discussing the difficulties that a learner of English can encounter in mastering easily confused words, French writes: "It is sometimes possible to discover a thread which will lead us [native speakers, instructors], and perhaps our pupils, through the maze" (French 1949:96). The aim of this thesis has been to discover such threads:

the creation of models that can help students, NLP applications, and even native speakers find their way through the maze of prepositions and determiners.

Bibliography

Luiz Amaral and Detmar Meurers. Where does ICALL fit into foreign language teaching? CALICO Conference presentation, www.ling.ohio-state.edu/icall/handouts/calico06-amaral-meurers.pdf, 2006.

Øistein Andersen. 2007. Grammatical error detection using corpora and supervised learning. In *Proceedings of the Twelfth ESSLLI Student Session*.

Guy Aston, Silvia Bernardini, and Dominic Stewart, editors. 2004. *Corpora and Language Learners*. John Benjamins, Amsterdam.

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater v.2. *Journal of Technology, Learning, and Assessment*, 4(3).

Eric Atwell. 1987. How to detect grammatical errors in a text without parsing it. In *Proceedings of the Third Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 38–45, Copenhagen, Denmark.

Roger Berry. 1993. *Collins Cobuild English Guides 3: Articles*. HarperCollins.

Johnny Bigert. 2004. Probabilistic detection of context-sensitive spelling errors. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC)*, Lisbon, Portugal.

John Bitchener, Stuart Young, and Denise Cameron. 2005. The effect of different types of corrective feedback on ESL student writing. *Journal of Second Language Writing*, 14:191–205.

Robert Blake. 2007. New trends in using technology in the language curriculum. *Annual Review of Applied Linguistics*, 27:76–97.

Thorsten Brants and Alex Franz. 2006. *Web 1T 5-gram Version 1*. Linguistic Data Consortium, Philadelphia.

Ted Briscoe, John Carroll, and Rebecca Watson. 2006. The second release of the RASP system. In *Proceedings of the COLING-ACL 06 Demonstrations Session*, pages 77–80, Sydney, Australia.

Chris Brockett, Bill Dolan, and Michael Gamon. 2006. Correcting ESL errors using phrasal SMT techniques. In *Proceedings of the Joint Conference of the International Committee on Computational Linguistics (COLING) and the Association for Computational Linguistics (ACL)*, pages 249–256, Sydney, Australia.

Lou Burnard, editor. 2000. *The British National Corpus Users Reference Guide*. British National Corpus Consortium, Oxford University Computing Services.

Jill Burstein, Martin Chodorow, and Claudia Leacock. 2003. Criterion online essay evaluation: an application for automated evaluation of student essays. In *Proceedings of the Fifteenth Annual Conference on Innovative Applications of Artificial Intelligence*, pages 3–10, Acapulco, Mexico.

Angela Chambers. 2005. Integrating corpus consultation in language studies. *Language Learning and Technology*, 9(2):111–125.

Jean Chandler. 2003. The efficacy of various kinds of error feedback for improvement in the accuracy and fluency of L2 student writing. *Journal of Second Language Writing*, 12:267–296.

Carol Chapelle. 2001. *Computer Applications in Second Language Acquisition: foundations for teaching, testing and research*. Cambridge University Press, Cambridge.

Carol Chapelle. 2007. Technology and second language acquistion. *Annual Review of Applied Linguistics*, 27:98–114.

Eugene Charniak. 2000. A maximum-entropy inspired parser. In *Proceedings of the First Conference of the North American chapter of the Association for Computational Linguistics (NAACL)*, pages 132–139, Seattle, Washington.

Nitesh Chawla, Nathalie Japkowicz, and Aleksander Kolcz. 2004. Editorial: Special issues on learning from imbalanced data sets. *SIGKDD Explorations*, 6(1):1–6.

Martin Chodorow and Claudia Leacock. 2000. An unsupervised method for detecting grammatical errors. In *Proceedings of NAACL*, pages 140–147, Seattle, Washington.

Martin Chodorow, Joel Tetreault, and Na-Rae Han. 2007. Detection of grammatical errors involving prepositions. In *Proceedings of the 4th ACL-SIGSEM Workshop on Prepositions*, Prague, Czech Republic.

Stephen Clark and James Curran. 2007. Wide-coverage efficient statistical parsing with CCG and log-linear models. *Computational Linguistics*, 33(4):493–552.

Michael Collins. 1999. *Head-Driven Statistical Models for Natural Language Parsing*. PhD thesis, University of Pennsylvania.

Stephen Pit Corder. 1981. *Error analysis and interlanguage*. Oxford University Press, Oxford.

Anthony Cowie, editor. 1998. *Phraseology: theory, analysis, and applications*. Oxford University Press, Oxford.

James Curran and Stephen Clark. 2003a. Investigating GIS and smoothing for maximum entropy taggers. In *Proceedings of EACL*, pages 91–98, Budapest, Hungary.

James Curran and Stephen Clark. 2003b. Language independent NER using a maximum entropy tagger. In *Proceedings of the Fourth Conference on Computational Natural Language Learning (CoNLL)*, pages 164–167, Edmonton, Canada.

James Curran, Stephen Clark, and Johan Bos. 2007. Linguistically motivated large-scale NLP with C&C and Boxer. In *Proceedings of the ACL 2007 Demonstrations Session*, pages 29–32, Prague, Czech Republic.

Walter Daelemans, Antal Van den Bosch, and Jakub Zavrel. 1999. Forgetting exceptions is harmful in language learning. *Machine Learning*, 34:11–41.

Rachele De Felice and Stephen Pulman. 2007. Automatically acquiring models of preposition use. In *Proceedings of the 4th ACL-SIGSEM Workshop on Prepositions*, pages 45–50, Prague, Czech Republic.

Pieter de Haan. 2000. Tagging non-native English with the TOSCA-ICLE tagger. In Christian Mair and Marianne Hundt, editors, *Corpus Linguistics and Linguistic Theory*, pages 69–79. Rodopi, Amsterdam.

Pieter de Haan and Hans van Halteren. 1997. *The TOSCA-ICLE Tagset - Software Manual*. TOSCA Research Group, University of Nijmegen.

Inge de Mönnink. 2000. Parsing a learner corpus? In Christian Mair and Marianne Hundt, editors, *Corpus Linguistics and Linguistic Theory*, pages 81–90. Rodopi, Amsterdam.

Marina Dodigovic. 2005. *Artificial Intelligence in Second Language Learning: raising error awareness*. Multilingual Matters, Clevedon.

Dan Douglas and Volker Hegelheimer. 2007. Assessing language using computer technology. *Annual Review of Applied Linguistics*, 27:115–132.

Jens Eeg-Olofsson and Ola Knutsson. 2003. Automatic grammar checking for second language learners – the use of prepositions. In *Proceedings of the Fourteenth Nordic Conference of Computational Linguistics (NODALIDA)*, Reykjavik, Iceland.

Rod Ellis and Gary Barkhuizen. 2005. *Analysing Learner Language*. Oxford University Press, Oxford.

Christiane Fellbaum, editor. 1998. *WordNet: an Electronic Lexical Database*. MIT Press, Cambridge, MA.

Dana Ferris. 2004. The grammar correction debate in L2 writing: Where are we, and where do we go from here? (and what do we do in the meantime ?). *Journal of Second Language Writing*, 13(1):49–62.

Jennifer Foster. 2004. *Good reasons for noting bad grammar: empirical investigations into the parsing of ungrammatical written English*. PhD thesis, Trinity College Dublin.

Jennifer Foster and Carl Vogel. 2004. Parsing ill-formed text using an error grammar. *Artificial Intelligence Review*, 21:269–291.

F.G. French. 1949. *Common errors in English: their cause, prevention, and cure*. Oxford University Press, London.

Yoav Freund and Robert Schapire. 1999. Large margin classification using the perceptron algorithm. *Machine Learning*, 37:277–296.

Michael Gamon, J. Gao, C. Brockett, A. Klementiev, W. Dolan, D. Belenko, and L. Vanderwende. 2008. Using contextual speller techniques and language modeling for ESL error correction. In *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)*.

Andrew Golding. 1995. A Bayesian hybrid for context-sensitive spelling correction. In *Proceedings of the Third Workshop on Very Large Corpora*, pages 39–53, Cambridge, MA.

David Graff. 2003. *English Gigaword*. Linguistic Data Consortium, Philadelphia.

Sylviane Granger. 1998a. The computer learner corpus: a versatile new source of data for SLA research. In Sylviane Granger, editor, *Learner English on computer*, pages 3–18.

Sylviane Granger, editor. 1998b. *Learner English on computer*. Longman, London.

Sylviane Granger. 1998c. Prefabricated patterns in advanced EFL writing: collocations and formulae. In Anthony Cowie, editor, *Phraseology: theory, analysis and applications*, pages 145–160.

Sylviane Granger. 2004. Computer learner corpus research: Current status and future prospect. In Ulla Connor and Thomas Upton, editors, *Applied corpus linguistics: a multidimensional perspective*, pages 123–145. Rodopi, Amsterdam.

Sylviane Granger, Estelle Dagneaux, and Fanny Meunier. 2002. *The International Corpus of Learner English. Handbook and CD-ROM*. Presses Universitaires de Louvain, Louvain-la-Neuve.

Sylviane Granger, Olivier Kraif, Claude Ponton, Georges Antoniadis, and Virginie Zampa. 2007. Integrating learner corpora and Natural Language Processing. *ReCALL*, 19(3):252–268.

Sylviane Granger and Fanny Meunier, editors. 2008. *Phraseology: an interdisciplinary perspective*. John Benjamins, Amsterdam.

Sidney Greenbaum and Randolph Quirk. 1990. *A Student's Grammar of the English Language*. Longman, London.

Ebba Gustavii. 2005. Target language preposition selection – an experiment with transformation-based learning and aligned bilingual data. In *Proceedings of the European Association for Machine Translation Conference (EAMT)*, pages 112–118, Budapest, Hungary.

Na-Rae Han, Martin Chodorow, and Claudia Leacock. 2006. Detecting errors in English article usage by non-native speakers. *Natural Language Engineering*, 12 (1):115–129.

Sylvana Sofkova Hashemi, Robin Cooper, and Robert Andersson. 2003. Positive grammar checking: a finite state approach. In *Proceedings of CICLing 2003*, pages 635–646.

Thomas Herbst, D. Heath, I.F. Roe, and D. Gotz. 2004. *A Valency Dictionary of English: a corpus-based analysis of English verbs, nouns and adjectives.* Mouton de Gruyter, Berlin/New York.

Archibald Hill. 1966. A re-examination of the English articles. In *Report of the Seventeenth Annual Round Table Meeting on Linguistics and Language Studies*, pages 217–231. Georgetown University Press, Washington, D.C.

Albert Hornby, editor. 1974. *Oxford Advanced Learner's Dictionary of Current English.* Oxford University Press, London.

Rodney Huddleston and Geoffrey Pullum. 2005. *A Student's Introduction to English Grammar.* Cambridge University Press, Cambridge.

Susan Hunston and Gill Francis. 2000. *Pattern Grammar: a corpus-driven approach to the lexical grammar of English.* John Benjamins, Amsterdam.

Emi Izumi, Kiyotaka Uchimoto, and Hitoshi Isahara. 2004. SST speech corpus of Japanese learners' English and automatic detection of learners' errors. *ICAME*, 28: 31–48.

Carl James. 1998. *Errors in Language Learning and Use: exploring error analysis.* Longman, London.

Nathalie Japkowicz. 2000. Learning from imbalanced data sets: a comparison of various strategies. In *Learning from imbalanced data sets: papers from the AAAI workshop*, Menlo Park, California.

Tuomo Kakkonen. 2007. Robustness evaluation of two CCG, a PCFG and a Link Grammar parsers. In *Proceedings of the 3rd Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, Poznan, Poland.

Evelien Keizer. 2004. Postnominal PP complements and modifiers: a cognitive distinction. *English Language and Linguistics*, 8(2):323 – 350.

Dan Klein and Christopher Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of ACL*, pages 423–430.

Kevin Knight and Ishwar Chander. 1994. Automated postediting of documents. In *Proceedings of the Twelfth AAAI Conference on Artificial Intelligence (AAAI)*, pages 779–784.

Miroslav Kubat and Stan Matwin. 1997. Addressing the curse of imbalanced training sets: one-sided selection. In *Fourth International Conference on Machine Learning (ICML)*, pages 179–186.

Mirella Lapata and Frank Keller. 2005. Web-based models for Natural Language Processing. *ACM Transactions on Speech and Language Processing*, 2(1):1–31.

Icy Lee. 2004a. Error correction in L2 secondary writing classrooms: the case of Hong Kong. *Journal of Second Language Writing*, 13:285 – 312.

John Lee. 2004b. Automatic article restoration. In *Proceedings of the NAACL Student Research Workshop*.

John Lee and Ola Knutsson. 2008. The role of PP attachment in preposition generation. In *Proceedings of CICLing 2008*, pages 643–658.

John Lee and Stephanie Seneff. 2006. Automatic grammar correction for second-language learners. In *Proceedings of Interspeech*, pages 1978–1981.

Seth Lindstromberg. 1995. *English prepositions explained*. John Benjamins, Amsterdam.

Charles Ling and Chenghui Li. 1998. Data mining for direct marketing: problems and solutions. In *Fourth International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 73–79.

Marcus Maloof. 2003. Learning when data sets are imbalanced and when costs are unequal and unknown. In *Proceedings of the Second Workshop on Learning from Imbalanced Datasets*, Washington DC.

Lidia Mangu and Eric Brill. 1997. Automatic rule acquisition for spelling correction. In *Proceedings of the International Conference on Machine Learning*, pages 734–741.

Anna Mauranen. 2004. Speech corpora in the classroom. In Guy Aston, Silvia Bernardini, and Dominic Stewart, editors, *Corpora and Language Learners*, pages 195–211.

Fanny Meunier. 1998. Computer tools for the analysis of learner corpora. In Sylviane Granger, editor, *Learner English on computer*, pages 19–37.

Guido Minnen, Francis Bond, and Ann Copestake. 2000. Memory-based learning for article generation. In *Proceedings of CoNLL*, Lisbon, Portugal.

Guido Minnen, John Carroll, and Darren Pearce. 2001. Applied morphological processing of English. *Natural Language Engineering*, 7(3):207–223.

Roger Mitton. A description of a computer-usable dictionary file based on the Oxford Advanced Learner's Dictionary of Current English. Online at ftp://ota.ox.ac.uk/pub/ota/public/dicts/710/text710.doc, 1992.

Ryo Nagata, Tatsuya Iguichi, Fumito Masui, Atsuo Kawai, and Naoki Isu. 2005a. A statistical model based on the three head words for detecting article errors. *IEICE Transactions on Information and Systems*, E88-D(7):1700–1706.

Ryo Nagata, Tatsuya Iguichi, Kenta Wakidera, Fumito Masui, Atsuo Kawai, and Naoki Isu. 2005b. Recognizing article errors in the writing of Japanese learners of English. *Systems and Computers in Japan*, 36(7):54–63.

Ryo Nagata, Tatsuya Iguichi, Kenta Wakidera, Fumito Masui, Atsuo Kawai, and Naoki Isu. 2006a. Recognizing article errors using prepositional information. *Systems and Computers in Japan*, 37(12):17–26.

Ryo Nagata, Atsuo Kawai, Koichiro Morihiro, and Naoki Isu. 2006b. A feedback-augmented method for detecting errors in the writing of learners of English. In *Proceedings of ACL-COLING*, pages 241–248, Sydney, Australia.

Diane Nicholls. 2003. The Cambridge Learner Corpus – error coding and analysis for lexicography. In *Proceedings of the Corpus Linguistics 2003 Conference*, pages 572–581.

Martin Parrott. 2000. *Grammar for English Language Teachers*. Cambridge University Press.

Maria Teresa Prat Zagrebelsky, editor. 2004. *Computer Learner Corpora – Theoretical issues and empirical case studies of Italian advanced EFL learners' language*. Edizioni dell'Orso, Alessandria.

Norma Pravec. 2002. Survey of learner corpora. *ICAME*, 26:81–114.

Adwait Ratnaparkhi. 1998. *Maximum Entropy Models for natural language ambiguity resolution*. PhD thesis, University of Pennsylvania.

Bertus van Rooy and Lande Schäfer. 2003. An evaluation of three POS taggers for the tagging of the Tswana Learner English Corpus. In *Proceedings of the Corpus Linguistics 2003 Conference*, pages 835–844.

John Saeed. 2003. *Semantics*. Blackwell, Oxford.

Patrick Saint-Dizier. 2005. Prepnet: a framework for describing prepositions: preliminary investigation results. In *Proceeding of the Sixth International Workshop on Computational Semantics (IWCS 6)*, pages 145–157, Tilburg, Netherlands.

Tony Silva and Melinda Reichelt. 2003. Second language writing up close and personal. In Barbara Kroll, editor, *Exploring the dynamics of second language writing*, pages 93–114. Cambridge University Press, Cambridge.

John Sinclair. 1991a. *Collins Cobuild English Guides 1: Prepositions*. HarperCollins.

John Sinclair. 1991b. *Corpus, Concordance, Collocation*. Oxford University Press, Oxford.

Jonas Sjöbergh. 2005. Chunking: an unsupervised method to find errors in text. In *Proceedings of NODALIDA*, pages 180–185, Joensuu, Finland.

Joel Tetreault and Martin Chodorow. 2008a. Native judgments of non-native usage: experiments in preposition error detection. In *Proceedings of the Coling 2008 Workshop on Human Judgements in Computational Linguistics*, pages 24–32, Manchester, UK.

Joel Tetreault and Martin Chodorow. 2008b. The ups and downs of preposition error detection in ESL writing. In *Proceedings of COLING*, pages 865–872, Manchester, UK.

James Thomas. 2004. Using computers in correcting written work. *Teaching English with Technology*, 4(3). URL http://www.iatefl.org.pl/call/j_soft18.htm.

Jesse Tseng. 2000. *The representation and selection of prepositions.* PhD thesis, University of Edinburgh.

Jenine Turner and Eugen Charniak. 2007. Language modeling for determiner selection. In *NAACL-HLT Companion volume.*

Andrea Tyler and Vyvyan Evans. 2003. *The Semantics of English Prepositions: spatial scenes, embodied meaning, and cognition.* Cambridge University Press, Cambridge.

Joachim Wagner, Jennifer Foster, and Josef van Genabith. 2007. A comparative evaluation of deep and shallow approaches to the automatic detection of common grammatical errors. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 112–121.

David Wible, C. H. Kuo, N. L. Tsao, A. Liu, and H. L. Lin. 2003. Bootstrapping in a language learning environment. *Journal of Computer Assisted Learning*, 19(1): 90–102.

Frederick Wood. 1967. *English Prepositional Idioms.* Macmillan, London.

Xing Yi, Jianfeng Gao, and William Dolan. 2008. A web-based English proofing system for ESL users. In *Proceedings of IJCNLP.*