

Research Statement
C. Paul Cook

My research in computational linguistics is primarily concerned with *lexical acquisition*, that is, automatically learning various properties of words, typically aspects of their syntax or semantics. I take a very interdisciplinary approach in my research. My research has shown that insights from linguistic theory can be effectively incorporated in statistical computational models of language; this stands in contrast to the currently popular trend in computational linguistics to rely primarily on statistical models, and much less so on linguistic theory. Furthermore, my research has shown that by exploiting linguistic information, the need for manually-annotated training data can be reduced or, in some cases, eliminated. This is important as such data is expensive to create, unavailable for many interesting problems, and furthermore, even when it is available, there is almost always never enough of it. My research also tends to concentrate on linguistic phenomena that go largely unnoticed in computational linguistics, but are nevertheless frequent and deserving of more attention. In this vein my research draws attention to important under-studied problems. To date my research has focused on two topics: neologisms and multiword expressions.

Automatically learning properties of new words Neologisms, or new words, are constantly being coined. Systems for natural language processing (NLP) tasks, such as machine translation or automatic question answering, often depend on lexicons for a variety of information, such as a word's parts-of-speech or meaning representation. Therefore, when a neologism is encountered in a text being processed, the performance of the entire system will likely suffer due to missing lexical information. Therefore, it is essential that an NLP system be able to identify neologisms as such, and infer various aspects of their syntactic or semantic properties necessary for the computational task at hand.

My research has considered automatically inferring properties of two common types of new word, namely lexical blends and the creative forms in text messaging. Lexical blends—words such as *cosmeceutical*, a combination of *cosmetic* and *pharmaceutical*—are a frequent type of new word. The similarly creative forms found in text messaging—such as *2nite* for *tonight*—are increasingly common, even outside of computer-mediated communication. Nevertheless, both lexical blends and text messaging forms have received little attention in computational linguistics. These word types are similar in that they are both *subtractive* word formations, that is, they are formed from partial material (characters and sounds) from existing words. In order to be able to understand expressions of these types, the reader or listener must be able to recover the underlying word or words. My work on lexical blends is the first computational study of this phenomenon. My research has focused on

inferring the source words of these expressions, e.g., determining that *cosmeceutical* is formed from *cosmetic* and *pharmaceutical*. Source word identification is an important first step in the automatic interpretation of lexical blends. I have also proposed preliminary methods for identifying lexical blends, i.e., distinguishing blends from other word types. My research on text messaging has considered the task of normalization, e.g., inferring the standard form *tonight* from the text messaging form *2nite*. This is an important problem that must be solved before other NLP tasks, such as machine translation, can be done effectively. I proposed an unsupervised model for text message normalization that exploited knowledge of common ways in which text messaging forms are created; this model performed as well as a supervised method for the same task. Moreover, unsupervised approaches to normalization can be adapted to other genres which also contain many non-standard forms—such as *tweets* from the micro-blogging service Twitter (<http://twitter.com>)—without the cost of developing gold-standard training data, unlike supervised approaches to this task.

Multiword expressions My second line of research has focused on multiword expressions (MWEs)—sequences of words which have an idiosyncratic interpretation—e.g., *hard drive*, *have a nap*, and *draw the line*. Such expressions are both frequent within a given language and common cross-lingually, and moreover, pose tremendous challenges for NLP. For example, systems must be able to distinguish between literal combinations, e.g., *She drew the line using red chalk*, and MWEs, e.g., *She drew the line at two hundred guests, and didn't invite the Smith's*. In collaboration with Afsaneh Fazly, formerly a postdoctoral fellow at the University of Toronto, I developed methods for distinguishing literal and idiomatic usages of idioms formed from a verb and a noun—as in the previous examples with *drew the line*—a very common type of English idiom. Drawing on linguistic knowledge of this class of idioms, we developed an unsupervised method for this task; such approaches are particularly important in this case given that gold-standard training data for many similar idioms is not available, and is required for supervised methods. The data from this study was made freely available for use by other researchers, and has since been used in a number of other studies of idioms.

My research on MWEs has also examined verb–particle constructions (VPCs)—e.g., *fill in*, *set up*, and *freak out*; in particular, I have considered automatically determining the semantic contribution of the particle (e.g., *in*, *up*, or *out*) in these constructions. Drawing on cognitive linguistic analyses of the semantics of VPCs, I developed a supervised method for classifying VPCs according to the meaning of their particle. For example, *up* has a different meaning in each of the following: *he heated up the soup*, *the cab drew up to the curb*, and *she cleaned up her room*.

Although numerous previous studies had considered the compositionality of such constructions—i.e., whether the verb or particle contribute their meaning to the overall expression—the issue of determining the meaning contributed by the particle had not been previously considered in computational linguistics.

Future research My future research will continue to consider problems related to neologisms and MWEs. Although new words are constantly being coined, established words also undergo changes in meaning or acquire new senses. For example, in Old English *meat* referred to food in general, but now has the more specific meaning of the flesh of animals. More recently, *bad*, *sick*, and *wicked* have all acquired new senses meaning roughly ‘good’. Identifying such changes in meaning is one of the most challenging problems in both lexical acquisition and lexicography. Many new MWEs are also created, and, like single words, these terms must be identified and added to lexicons; indeed, a quick glance at a dictionary documenting new words—e.g., the Double-Tongued Dictionary (<http://www.doubletongued.org>)—reveals many recently-coined MWEs. Identifying new MWEs is especially challenging given that MWEs may have the same surface form as a literal combination of words, which typically should not be entered into a dictionary. These specific projects are part of my longterm research objective, which is to develop computational methods to automatically, or semi-automatically, identify and document language change, and differences between speech communities defined by variables such as geographical location, age, or socio-economic status. Given the huge amount of text available on the World Wide Web, and the wide range of speech communities it represents, along with the fact that language is constantly changing, computational linguistics as a community must seriously address problems related to language variation and change in order to succeed in producing high-quality, domain and genre independent, NLP systems.