the Penguin

Using the Web as a Database for Descriptive and Dynamic Grammar and Spell Checking

In consequence of emergent limitations of traditional spell and grammar checkers, the Penguin prototype system has been designed to be a descriptive and dynamic tool for assisting in computer based writing. Rather than relying on a static dictionary, the World Wide Web is used as a database to handle language artifacts out of the ordinary, such as idioms, colloquialisms, names, and slang expressions; a common source of concern especially for second language speakers

Introduction

In linguistics, prescriptivists and descriptivists argue about grammar and spelling. Prescriptive grammarians claim that it's possible to say what language should be like, while their descriptive colleagues hold that we should only describe language, not impose right and wrong on people.

But this disagreement needs to be examined in the light of two widely agreed upon characteristics of language. First, all languages have variation, e.g. in terms of differences between regions, social class, ethnicity, level of formality, age, and gender. Second, changes in language affecting pronunciation, vocabulary, and grammar appear over time. Hence, change is inevitable, even to the prescriptivists.

In computing, dictionary based spell and grammar checkers (typically built into word processing applications) tend to be prescriptive, e.g. building on models of what is correct use of grammar. They are static, too, as the rules and dictionaries are locally stored and not regularly updated, and also notorious for not handling idioms, colloquialisms, names, and slang. In addition, traditional computer based spell checkers don't provide contextual information, such as how common a word is, or if a word is informal or esoteric; it's either in the dictionary or not.

Using the Web as a Database

To provide computer users with a creative tool to assist writing, specifically with the intent to provide descriptive spell and grammar checking, we have implemented a prototype system called *the Penguin,* which makes use of the World Wide Web in a novel way. The underlying idea is that web pages do not only contain information in that they focus on one or more topics; the words, sentences, word order and punctuation that constitute a particular web page is in fact also a vigorous representation of the language in which the page is written, i.e. the whole of the web contains information that transcends the particular subject matter of its individual documents. It is to consider the web a huge, dynamic, non-normative and independent source of knowledge; in short, to think of the Web as a database.

The Penguin Prototype System

The basic operation the Penguin performs is to retrieve from the Web the number of occurrences of any given string—i.e. the number of instances real people have used a piece of language for real purposes—and present that figure to the user, who acts according to his or her interpretation of what that figure means.

Examples of Use

First, a user may be uncertain as to how *retrieve* is spelled (that is, by most people), but can imagine three likely alternatives. A query on *retrieve* returns 3,178,419 uses, while both *retreive* and *retrive* produce substantially less. While the Penguin indicates that one is more common, it does not state that one of the two alternatives is the only, correct spelling.

Second, the user knows that *hippo* is a common short for *hippopotamus*. But how common? And when should one use which? The Penguin suggests that while the phrase *hippo is* is more frequent than *hippopotamus is*, *hippopotamus is a large* is more common that *hippo is a large*, suggesting that *hippo* is indeed a common short name, but that in formal writing *hippopotamus* might still be preferred.

Third, a second-language English speaker wants to use the idiom *a bird in the hand*, but is not sure if it was that or *a hand in the bird?* The Penguin suggests the first to be much more common, and also signifies that it's probably an idiomatic expression, as the number of uses seem to constitute a critical mass not generally established by five-word combinations.

Discussion

We put forward the Penguin system primarily as a useful complement to traditional spell and grammar checkers. First, the Penguin is descriptive and non-normative, as it bases its suggestions on real use of language, not on prescriptive models of what language should be like. Second, it's dynamic in that the Web (its database) is constantly undergoing change, reflecting transformation of culture and society. The Penguin is thus able to handle new words as they appear. Third, it doesn't judge language subjectively in that no words that are in use are excluded. Fourth, as the Penguin does not embody a model of language, e.g. in terms of grammatical rules or systems of phonetics, it works with all languages represented on the web. Fifth, it adds user value as it doesn't just suggest whether or not a word exists, but also provides useful contextual information in terms of incidence. Sixth, the Penguin handles the idioms, colloquialisms, names, and slang expressions for which traditional spell checkers have no support.

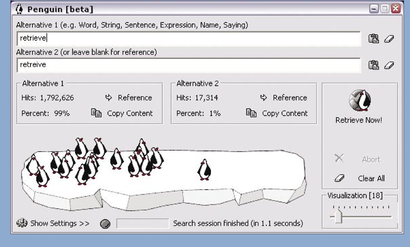
Future Work

The working hypothesis currently being tested is that the Penguin is especially useful for people writing in a foreign language, where certain language artifacts do not come naturally. As a further development, we also plan to add context definitions to partition the Web as a database, primarily based on time, e.g. between 93-95; place, e.g. restricted to *.gov* domains; and subject, e.g. how people interested in witchcraft spell *cast*.

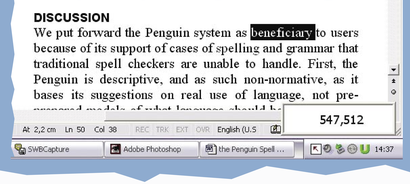
Daniel Fallman
Interactive Institute.se

Tools for Creativity Studio
Umeå, Sweden
daniel.fallman@interactiveinstitute.se

Normal Mode *lets the user compare alternative spellings or sentences, or compare alternatives with reference values, after which the results are visualized.*



Tray Mode *integrates the Penguin transparently with other applications and mediates the result to the user through an unobtrusive pop up message window. It offers users instant feedback without the need for switching applications.*



http://penguin.fallman.org
Patent Pending