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

Daniel Fallman
Interactive Institute, Tools for Creativity Studio
Tvistevägen 47, Box 7914, 903 33 Umeå, Sweden
+46 90 18 51 46
daniel.fallman@interactiveinstitute.se

ABSTRACT
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 computer based writing. Rather than relying on a static dictionary, the 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.

Keywords
Penguin, grammar & spell checking, linguistics, prototype

INTRODUCTION
In the linguistics discipline, a prominent dispute is that between prescriptivists and descriptivists of grammar and spelling. At large, what separates these two groups is that prescriptive grammarians tend to ask the question *"what should language be like?"* while their descriptive colleagues ask *"what is language like?"* [3]. Although modern grammarians tend to describe rather than prescribe linguistic forms and their uses, prescriptive statements about wrong and right is nevertheless still common, e.g. in dictionaries such as *The New Fowler's* (see [2]). This disagreement, however, should be examined in the light of two widely agreed upon facts about language. First, all languages have variation, in terms of differences between e.g. regions, social class, ethnicity, level of formality, age, and gender [4]. Second, changes in language over time affect pronunciation, vocabulary, and grammar. Change is often seen as necessary, as language is part of culture and needs to be able to express its transformations [1].

Computerized Spell and Grammar Checking
In the world of computing, the dictionary based spell and grammar checkers built into applications like Microsoft Word tend to lean towards the prescriptive, as they are typically based on models of what a sentence *should* look like. As a rule, they are also static, as their dictionaries are stored locally and as these files are not regularly updated. Furthermore, traditional spell checkers are notorious for not handling expressions such as idioms, colloquialism, names, and slang. These all represent a highly dynamic character of language, out of reach of the prescriptive models on which traditional spell checkers rely. In addition, traditional computer based spell checkers do not provide contextual information such as how common a word is, or if a word is informal or esoteric, they merely imply whether or not a word exists. Finally, they are also subjective in that some words for one reason or another are not included, and as such do not exist according to the traditional spell checker.

CHI 2002, April 20-25, 2002, Minneapolis, Minnesota, USA.
ACM 1-58113-454-1/02/0004.

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: Spell by the Web*, 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. The purpose here is to consider the web a huge, dynamic, non-normative and independent source of knowledge; or, 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 hits 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. The Penguin then lets the users act on the result according to their interpretation of what the results mean. The current Penguin client for Microsoft Windows has two basic modes of operation, *Normal Mode* and *Tray Mode*.

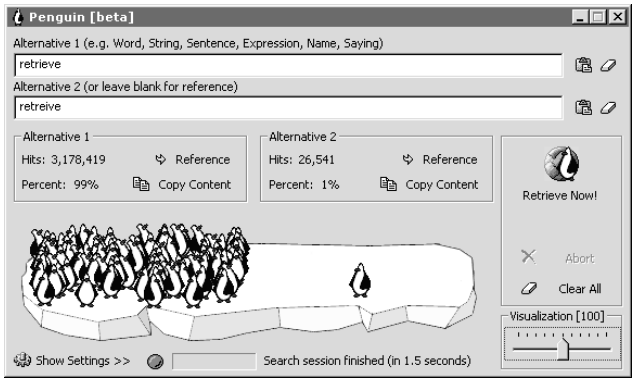

Figure 1. A screenshot of the Normal Mode

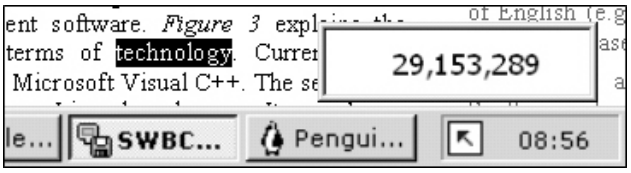

Figure 2. Screenshot of the Tray Mode

Normal Mode lets the user compare two or more alternative spellings or sentences, or compare one alternative with a reference value, after which the results are analyzed and visualized. Tray Mode integrates the Penguin transparently with other applications by watching the clipboard for changes and promptly mediates the result to the user through a pop up message window containing the number of hits for a given string, typically a selection from Microsoft Word. The Tray Mode hence offers users instant feedback without the need for switching applications.

Implementation

The Penguin is implemented as a Client/Server system, where the client sends the string to be searched for to the server, which accepts the string and connects to a search engine. The server queries the search engine, and collects the number of hits of the string, which is then returned to the user via the client software. *Figure 3* explains the Penguin system in terms of technology. The current client is developed in *Microsoft Visual C++*. The server is a *Perl 5* script running on a Linux based server. It provides an open interface, so developing clients for other platforms or with different front ends is a straightforward process.

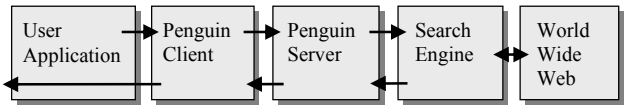

Figure 3. Penguin Prototype System Overview

THREE EXAMPLES OF USE

First, a user may be uncertain as to how *retrieve* is spelled, 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 just 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 a common short name, but that in more formal writing, hippopotamus is still to prefer.

Third, the second-language English speaker wants to use the idiom *a bird in the hand,* but is not sure if it was that or was it 'a hand in the bird?' The Penguin not only suggests the first to be much more common than the latter (actually 5,527 to 19), but also signifies that the first alternative is 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 as beneficiary to users because of its support of cases of spelling and grammar that traditional spell checkers are unable to handle. First, the Penguin is descriptive, and as such non-normative, as it bases its suggestions on real use of language, not pre-prepared models of what language should be like. Second, it is dynamic in that the web is constantly undergoing change. Hence, the Penguin's database is automatically updated in real time and as such handles new words as soon as they appear, and will hence be able to follow changes in language dynamically. Third, it does not judge language subjectively in that no words that are in use are excluded, and is thus a very rich depiction of language. Fourth, as the Penguin does not embody a model of language, e.g. in terms of grammatical rules or a system of phonetics, it works with all languages represented on the web. Fifth, it adds user value as it does not just suggest whether or not a word exists, but it also provides useful contextual information in terms of incidence. Sixth, as indicated previously, the Penguin handles the idioms, colloquialisms, names, and slang expressions for which traditional spell checkers have no support. In addition, and as a pertinent HCI issue, the user is endowed with a certain amount of ambiguity at the interface: what does a result like 5,527 really mean? As the Penguin does not suggest or judge, it depends and relies on the user to interpret the results.

FUTURE WORK

The working hypothesis 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. how 'nerd' is spelled at MIT; and subject, e.g. how people interested in witchcraft spell 'cast'. At present, the Penguin is subject to extensive long-term testing, in which a large focus group employ it on a day-to-day basis. Log files of all queries are being collected to help us understand if use of the Penguin differs significantly from that of a traditional spell checker.

CONCLUSIONS

The Penguin is a descriptive, non-normative, dynamic, and language independent tool for spell and grammar checking that uses the web as a database. It should also be seen a useful compliment to traditional spell checkers as it handles language artifacts such as idioms, colloquialisms, names, and slang expressions.

REFERENCES

1. Barber, C., *The English Language,* Cambridge University Press, Cambridge U.K., 1994
2. Fowler, H. W., *The New Fowler's Modern English Usage,* Oxford University Press, New York NY, 1996.
3. Fromkin, V. & Rodman, R., *An Introduction to Language,* Harcourt Brace, New York NY, 1993
4. Wolfram, W., *Dialects and American English,* Prentice Hall, Englewood Cliffs NJ, 1991