# The Typology of Unknown Words: An Experimental Study of Two Corpora

**Xiaobo Ren and François Perrault**

xren@ccrit.doc.ca, perrault@ccrit.doc.ca
CCRIT, Communications Canada, 1575 Chomedey Bld, Laval, Québec, Canada, H7V 2X2

## Table of contents



## 1.0 Introduction

Most current state-of-the-art natural language processing (NLP) systems, when presented with real-life texts, have problems recognizing each and every word present in the input. Depending on the application, the consequences can be severe. For example, in a machine translation system the quality of the processing may suffer and sometimes further processing may even be impossible.There are two main reasons why a word might not be recognized and thus be considered *unknown* by the system:

- The linguistic knowledge of the system is not complete, i.e. the word is correct but is not present in the system's dictionary;
- The word is erroneous.

A lot of effort has been directed towards dealing with the latter, i.e. finding ways of detecting and correcting erroneous words. Most of the developments in this area of research are based on a paper by Damerau [Damerau 64] where the author offers a classification of erroneous words.

The aim of this paper is to present further results about the frequency and types of unknown words found in real-life corpora. We hope that the results of our study will be of some use in the development of NLP systems capable of dealing with realistic input.

Our findings confirm Damerau's results in that the great majority of erroneous words contain a single typographical error and belong to one of the four following categories: insertion, deletion, substitution, transposition.

But we have also found that a large proportion of the unknown words is made up of correct words which are not present in the dictionary. For example, derived words alone represent 30% of all unknown words in our samples.

These results indicate the need for further work before an acceptable level of robustness can be attained. Although traditional typographical error detection and correction techniques can be used to handle the majority of erroneous words, much remains to be done before such problematic areas as derived words can be dealt with effectively.

## 2.0 Related work

In his pioneering article [Damerau 64], the author gives valuable information about the frequency of typographical errors. In his paper Damerau indicates that typically, 80% of all ill-formed words in a document are the result of one of four typographical errors:

- Transposition of two letters, e.g. étbali instead of établi;
- Insertion of one extra letter, e.g. économioque instead of économique;
- Deletion of one letter, e.g. additionelle instead of additionnelle;
- Substitution of a valid letter by one that is wrong, e.g. ogligé instead of obligé.

More recent results [Pollock and Zamora 83] also indicate that in most cases, there is only one error per word.

The classification of possible errors has been extended over the years to include other types of errors [Srihari 85, Szanzer 69, Veronis 88]. Based on this body of work, we

can propose the following incomplete list of the possible nature of errors:

- Typographical errors, which are errors of *execution* in carrying out the task of typing text on a keyboard;
- Orthographic errors, which are errors of *intention* attributable to distraction or lack of knowledge on the part of the author;
- Syntactic and semantic errors;
- Errors committed during the input procedure, either by an optical character recognition device or by a speech recognition system;
- Storage and transmission errors due to noisy electronics or communication channels.

## 3.0 Corpus

Our typology of unknown words is based on the study of two corpora.

### 3.1 Hansard

The first one, a French corpus called the Hansard, is a transcript of all the proceedings that took place in the Canadian House of Commons in 1986.

Since Canada is officially a bilingual country, whenever Members of Parliament gather together to debate laws, the transcripts of the session have to be made available in both English and French. On the day a session is held, transcripts are translated and printed rapidly in order for the Members of Parliament to have a bilingual copy of the previous days' session on their desk the next morning.

The main characteristics of this corpus are:

- Spoken language style;
- Manually typed on a computer;
- Made up of both translations from English to French and source French statements;
- Translated by qualified professional translators;
- Translated rapidly;
- Even the source text is sometimes touched up by professional writers.

### 3.2 Jobs

The second corpus, called Jobs, was obtained from Employment and Immigration Canada and consists of English job offers. Employment centres across Canada receive calls from employers offering job opportunities. Clerks are responsible for answering the telephone and writing up the job postings.

The main characteristics of this corpus are:

- Telegraphic style;
- Manually typed into a computer program that has a rigidly formatted interface;
- Made up solely of text originally written in English;
- Written rapidly by a clerk.

### 3.3 Extracting unknown words

The two corpora differ in nature and in the way respective lists of unknown words were extracted.

For the Hansard corpus we tokenized the text and we automatically tagged each token with a part of speech [Foster 91]. From this list we then removed all punctuation, numbers and words beginning with a capital letter (proper nouns and abbreviations merit separate study). We then singled out all the words that could not be found in an electronic dictionary. For this operation we used the DMF [Bourbeau, Pinard 86] which contains the equivalent of 59 000 entries.

As for the English corpus most of the work was done by hand. We tokenized the text as previously described but the sifting of punctuation, numbers, words beginning with a capital letter and known words, was done manually, leaving a list of unknown words.

## 4.0 Typology

We have divided the list of unknown words into two main groups. G1 contains words that could not be recognized but were correct, while G2 contains erroneous words. We have further subdivided these two groups into different types of unknown word.

Our goal has been to identify tendencies in this group we call "unknown words". In doing so, we increased the number of types and inevitably some of these types intersect. We have relied on our intuition and experience to assign the most plausible type to the unknown words.

In this section descriptions will be given of each of these types along with numerous examples. In addition, in the case of G2 types, we speculate on the possible causes of error.

### 4.1 G1: Correct words

#### 4.1.1 Proper nouns

In principle, proper nouns should not be part of the list of unknown words since we removed all words beginning with a capital letter. But a few occurrences of proper nouns appeared with the wrong capitalization and in other cases a lower case component of a proper noun (isolated by the tokenization process) was found.

E.g. ottawa (Ottawa)[1], nai (B'nai Brith)

#### 4.1.2 Abbreviations

Upper case abbreviations (acronyms, initials, etc.) are not considered to be unknown words, but a few (common) abbreviations are written in lower case and thus end up in the unknown word list.

E.g. km (kilomètre), pub (publicité)

#### 4.1.3 Ordinals

Although numbers and punctuation have not been considered valid unknown word candidates, since letters are sometimes used as roman numbers, a few ordinal numbers were found.

E.g. i (1), iv (4)

#### 4.1.4 Regional words

Those are words or expressions that cannot be found in traditional dictionaries. Some of them can found in specialized dictionaries [Shiaty 88] and some of them can be identified by native speakers.

E.g. abrier (couvrir), bécosses (toilettes), cenne (sou)

#### 4.1.5 Scholarly words

Scholarly words include technical or rare words. They can be found in large reference tools like Termium[2].

E.g. écosphère, amoxicillin, anadrome, ayatollah

#### 4.1.6 Parts of expressions

Certain expressions (French and Latin mostly) are made up of several elements separated by spaces. Isolated from the rest of the expression, some of these elements cannot be recognized.

E.g. facto (de facto), wa (oskee wa wa), feminem (ad feminem)

#### 4.1.7 Foreign words

In the Hansard this category corresponds to anglicisms or English words appearing in a quote.

E.g. abortionniste, affluente, runnés

However, we also found foreign words in the English corpus.

E.g. chao chow, noel, solicite

#### 4.1.8 Derived words

Derived words are very productive. The number of occurrences of this type of unknown word in the Hansard represents almost 30% of all unknown words. In French we found 96 affixes that were used to form new words. Certain words have both a prefix and a suffix at the same time.

E.g. **réé**chelonn**ement**, **pré**commercialis**ation**

Certain affixes are more productive than others:

anti-, dé-, dés-, extra-, sur-, in-, inter- ré-, super-
-age, -ation, -ien, -eur, -iser, -ment

#### 4.1.9 Compounds

We excluded from the unknown word list compounds beginning with a capital letter and compounds that cannot be recognized when considered as a whole nor when the elements are considered individually. The unknown words classified as compounds are: ones that should start with a capital letter but do not; those in which the necessary spaces or hyphens have been deleted, i.e. the elements have been concatenated; and compounds made up of an element that cannot be recognized (often because of the 'o' infix).

E.g. câblodistributeurs, chimio-dépendance, radioastronomique

#### 4.1.10 Garbled words

We include in this category words that are divided by a blank space, words that are joined together but are not compounds and words which are, in general, affected by electronic noise. Although in some ways this could be considered an error, we did not want to put this category in G2 because contrary to other types in G2, in this case the writer cannot be held responsible for the error.

E.g. employEs, sAvez-vous, afinque, erreur.Ce

### 4.2 G2: Erroneous words

#### 4.2.1 Accents

These errors are unique to the French corpus and can be subdivided into four types:

- Accent insertion.

  E.g. élèvant (élevant), éssai (essai), ôtages (otages)

- Accent deletion.

  E.g. achetera (achètera), aerospatiale (aérospatiale), agées (âgées)

- Substitution of one accent for another.

  E.g. àgées (âgées), évènement (événement), allègera (allégera)

- Repositioning of the accent.

  E.g. chomâge (chômage), composeé (composée), dégoûtant (dégoûtant)

#### 4.2.2 Punctuation

This type of error is unique to the English corpus and corresponds to problems with hyphens and apostrophes. There are three cases:

- Deletion of a necessary hyphen.

---

1. In the context of an example, parentheses indicate the correct or intended word.

2. The terminological data bank of the Translation Bureau of the Department of the Secretary of State of Canada.

E.g. cardio respiratory (cardio-respiratory), cleanup (clean-up)

- Insertion of a hyphen. This usually occurs when hyphens are used to parenthesize text.

  E.g. class-secondary, cleaners-including

- Deletion of an apostrophe denoting possession.

  E.g. companys (company's)

### 4.2.3 Insertions

Knowing the configuration of a standard keyboard and the way people type suggests several plausible reasons for the insertion of superfluous characters.

- A key is held down too long, generating sequences of identical letters.

  E.g. étonnne, accès, beaaucoup, paartnership

- The finger strikes two contiguous keys at the same time.

  E.g. économioque, égalememnt, profcessional, thgen

- 'Influence' of other letters in the same word.

  E.g. évidemement, aéoroport, accueuillir, taboubli, electrolologist

Other instances of insertion seem to be simply attributable to a lack of knowledge of the language.

E.g. éperduement, absoluement, orthopaedic, paediatric.

For another group of insertion-type errors, no obvious explanation could be found.

E.g. constinué, lotusi, manchine, experiencep

### 4.2.4 Deletions

The omission of a character is the most common typographical error. This is probably related to a situation where rapid typing is required and where the mind might work faster than the hand. Here is a list of the ten most frequently omitted letters (the percentages are based on the total number of words in this class):

| Letter | r | s | i | n | t | e | p | c | l | a |
|---|---|---|---|---|---|---|---|---|---|---|
| % | 9.4 | 9.2 | 6.3 | 5.8 | 5.8 | 5.4 | 3.2 | 3.1 | 2.5 | 2.2 |

**Table 1: Most common deletions in the Hansard**

| Letter | e | i | r | a | n | s | u | c | t | h |
|---|---|---|---|---|---|---|---|---|---|---|
| % | 9.2 | 6.4 | 5.6 | 4.5 | 4.1 | 3.8 | 3.0 | 2.6 | 2.6 | 1.9 |

**Table 2: Most common deletions in Jobs**

### 4.2.5 Substitutions

This is a fairly complex category. Substitution of one letter for another can be typographical or orthographic in nature. Some tentative explanations include:

- The letter is replaced by an adjacent letter.

  E.g. indident (incident), esperience (experience), satisfaisamte (satisfaisante)

- The wrong hand is used.

  E.g. quesque (quelque), gouvervement (gouvernement)

- The letter is 'influenced' by another letter in the same word.

  E.g. bubget (budget), songages (sondages), stell (steel)

- The error is orthographic in nature.

  E.g. maintenence (maintenance), engouragement (encouragement), nauvrage (naufrage)

- Other substitutions escape simple explanations.

  E.g. da (de), ja (je), saire (faire)

### 4.2.6 Transpositions

There are three types of transpositions:

- Inversion of adjacent letters.

  E.g. appearnace, appropraite, commerical

- Inversion of non-adjacent letters.

  E.g. énocomique, anamolic, condiser, ditues

- Although not strictly speaking a transposition, we also include here the displacement of a single letter.

  E.g. avatanges, comagpnies, avalaible, expierence

### 4.2.7 Grammar

There are not many errors under this heading, since no syntactic analysis has been done in order to extract the list of unknown words. What we have here are errors of morphology and conjugation.

E.g. étée (été), plues (plu), cloths (clothes)

### 4.2.8 Other

There are a few remaining words which we could not fit in the other categories; some of them are incorrect while others can be considered spelling variations that are not fully standard.

E.g. tee shirt (T-shirt), thru (through)

## 5.0 Frequency of unknown words

The Hansard corpus contains 4 173 506 tokens. Among these tokens we found 2 982 distinct unknown words occurring 9 301 times. This represents 0.2% of all tokens. The Jobs corpus contains 140 482 tokens. Of those, 1 016 were distinct unknown words occurring 2 109 times. This represents 1.5% of all tokens.

We now present in tabular form the frequency distribution of unknown words in both corpora. For each type of unknown word we indicate the number of distinct words (cases) and the total number of occurrences (occ.) found.

For each of these numbers, we also give the associated percentages over the *total* number of unknown words in both G1 and G2. Therefore the total percentages of G1 and G2 add up to 100 percent.

| Type | # of cases | % | # of occ. | % |
|---|---|---|---|---|
| Derived words | 526 | 17.22 | 2814 | 29.93 |
| Foreign words | 392 | 12.83 | 2014 | 21.42 |
| Scholarly words | 73 | 2.39 | 658 | 7.00 |
| Parts of expressions | 73 | 2.39 | 579 | 6.16 |
| Garbled words | 94 | 3.08 | 296 | 3.15 |
| Compounds | 48 | 1.57 | 160 | 1.70 |
| Ordinals | 8 | 0.26 | 153 | 1.63 |
| Abbreviations | 10 | 0.33 | 43 | 0.46 |
| Regional words | 18 | 0.59 | 26 | 0.28 |
| Proper nouns | 8 | 0.26 | 21 | 0.22 |
| Total | 1250 | 40.92 | 6764 | 71.93 |

**Table 3: G1 frequencies in the Hansard**

| Type | # of cases | % | # of occ. | % |
|---|---|---|---|---|
| Deletions | 645 | 21.11 | 976 | 10.38 |
| Insertions | 406 | 13.29 | 503 | 5.35 |
| Accents | 248 | 8.12 | 414 | 4.40 |
| Substitutions | 230 | 7.53 | 319 | 3.39 |
| Grammar | 141 | 4.62 | 258 | 2.74 |
| Transpositions | 135 | 4.42 | 169 | 1.80 |
| Total | 1805 | 59.08 | 2639 | 28.07 |

**Table 4: G2 frequencies in the Hansard**

| Type | # of cases | % | # of occ. | % |
|---|---|---|---|---|
| Garbled words | 143 | 13.64 | 322 | 15.27 |
| Foreign words | 10 | 0.95 | 36 | 1.71 |
| Total | 153 | 14.78 | 358 | 16.97 |

**Table 5: G1 frequencies in Jobs**

| Type | # of cases | % | # of occ. | % |
|---|---|---|---|---|
| Punctuations | 224 | 21.35 | 514 | 24.37 |
| Deletions | 287 | 27.36 | 467 | 22.14 |
| Substitutions | 158 | 15.06 | 363 | 17.21 |
| Insertions | 140 | 13.35 | 227 | 10.76 |
| Transpositions | 58 | 5.53 | 87 | 4.13 |
| Others | 13 | 1.24 | 49 | 2.32 |
| Grammar | 16 | 1.53 | 44 | 2.09 |
| Total | 894 | 85.22 | 1751 | 83.03 |

**Table 6: G2 frequencies in Jobs**

The following points should be noted:

- A word containing two errors is accounted for in two categories. This explains why the total is a slightly higher than the total number of unknown words given previously.
- In the Hansard there are 16 words (0.17%) that contain more than one error per word and 94 words (1.01%) that belong to both G1 and G2 (e.g. a word can be incorrect and be derived at the same time). On the other hand, with Jobs there are 42 words (1.99%) that contain more than one error per word. These results are comparable to Damerau's findings about the preponderance of single error words.
- Of course, different extraction procedures give different results. The Hansard contains a great many correct words not in the DMF; on the other hand the Jobs list of unknown words contains very few of those correct words. When faced with a word they do not recognize immediately, humans have the option of consulting a dictionary (general or specialized) and even if the word is not in any of those, the person can still rely on his or her intuition about word composition and derivation in order to accept a word.
- In the case of the Hansard the total number of occurrences in G1 (71.93%) is much higher than the total number of occurrences in G2 (28.07%). This significant result shows that instead of putting all of our efforts into trying to develop a better error corrector, we would gain a lot from looking into ways of dealing with the deficiencies of our lexical databases.
- Since English does not have accents, this category is not represented in G2 of Jobs.
- On the other hand, errors involving hyphens and apostrophes are very common in the Jobs corpus. We classified these as punctuation errors.
- We believe that the punctuation category of G2 Jobs is not representative of English in general. The high frequency of this type of error is due to a peculiarity of the

program responsible for the input of job descriptions which encourages the use of hyphens to parenthesize text.

## 6.0 Recognizing unknown words

In this section we examine possible avenues of investigation designed to deal with the different unknown word types.

### 6.1 G2: Erroneous words

When confronted with an unknown word, the ideal NLP system would be able to understand the text and to deal with what was intended by the writer, and not just what he wrote. But of course this is not within the scope of current technology.

A more realistic goal is to try to deal with typographical errors and a lot of attention over the years has been given to the detection and correction of such errors. Different methods have been proposed, some completely automatic, others meant to assist humans in proof reading, some practical and usable, others of theoretical interest only. For a good overview of this field of research we suggest [Peterson 80], while [Pollock 82] contains an extensive bibliography.

Despite years of research, the detection and correction of typographical errors remains a problem not entirely resolved. Commercial software as well as state-of-the-art techniques described in the literature can only propose approximate solutions. No program is capable of detecting *every* error and capable of always suggesting *the* right correction.

Despite their limitations, some existing methods can still be useful and sometimes even better than most human correctors. This fact is well illustrated by the success of commercial detector/correctors available on the market, despite an overall performance that can at best be described as acceptable [Dinnematin and Sanz 90].

In order to detect errors most techniques rely on a list of correct words known to the system (a dictionary), possibly augmented by a set of morphological rules.

Amongst the possible approaches to typographical error correction, two methods seem to be more successful than the others.We can either compare the unknown word against each of the dictionary words and if one of those comes close enough to the original word according to some measure of similarity, it can be used in its place (for an example see [Wagner and Fischer 74]). Or we can take an erroneous word, undo all possible errors we want to detect and then search the dictionary to see if any of those potential corrections produces a valid word. We call this method the hypothesis generation method. For example a transposition error can be detected by transposing each pair of characters in the unknown word and then consulting the dictionary with the resulting words. This essentially is the technique used in such programs as the DEC-10 Spell software.

The method based on a measure of similarity is too inefficient to be practical and is mostly of theoretical interest. The latter is more efficient but also more approximate in that it is not guaranteed that we will find a correction if we did not expect the offending error.

In both cases the contents of the dictionary must be carefully selected. It must be large enough to offer reasonable coverage, but on the other hand there is a real danger of using a list of words that is too big, in that a very extensive list will usually contain rare and archaic words that could correspond to errors on more frequent words.

An error corrector integrated in an NLP system should allow us to reduce the dictionary search space by comparing the erroneous word only with dictionary words complying with the syntactic and semantic requirements valid at that time in the processing. This should make the search significantly more efficient. For example, if at some point we are expecting a verb and we encounter an unknown word, in order to suggest corrections we could limit ourselves and consider only the verbs in the dictionary.

One interesting aspect of typographical error correction methods such as the hypothesis generation method is that they can also be used to correct some of the other types of errors. So with these methods, not only do we have a (somewhat approximate) solution to insertion, deletion, transposition and substitution errors, but in some cases they will also solve punctuation, accent and grammar errors. For example in the case of accents, we can extend the French alphabet with the possible accented letters and simply use this alphabet to generate more candidate corrections.

Again if the hypothesis generation method is chosen, then further use can be made of the knowledge gained about the type of errors usually committed. For example in order to minimize the number of hypotheses generated and to maximize the probability of finding the right correction, when testing the deletion of a character, one could attempt to "re-introduce" the character only in the case of the 10 most frequent deletions. More anecdotal knowledge gained through the sifting of the list of unknown words could also be of some use. For example, duplication of consonants was a frequent type of insertion error and thus, if only a few hypotheses are to be tried, unknown words with duplicate consonants could be considered prime candidates for insertion errors.

### 6.2 G1: Correct words

The results collected in the course of our study should at the very least, influence the amount of effort put into dealing with each of the different types of errors. The re-

alization that a large percentage of unknown words are part of the G1 group warrants renewed effort in treating this type of problem.

There will always be words that a system cannot recognize, if only because some of them belong to so-called open classes. But we can still reduce the number of such words.

One obvious solution is to enrich the dictionary, for example with common abbreviations and expressions. Another similar, but more modular solution consists in supplementing the basic dictionary with auxiliary dictionaries. One could envision separate dictionaries for regional and scholarly words for example.

The ordinals found in our corpora could easily be recognized by a grammar describing the formation of roman numerals.

Foreign words represent a difficult problem. They are exceptions to the usual assumption that the whole text to be processed is expressed in the same language throughout. Although it does not completely solve the problem, the detection of such signs as double quotes, setting the words apart from the text, could be used to suggest that the following unknown words might be foreign.

It might be possible to recognize garbled words and compounds by using methods similar to the ones used to treat G2 words. For example the deletion of a necessary hyphen could be detected and possibly corrected as is done for the deletion of an ordinary character.

As we have seen, derived words represent an impressive percentage of the total number of unknown words. Even if we were to enlarge the dictionary we would never be able to include every derived word, for they are much too productive. Therefore the solution seems to lie in a rule-based description of derivation similar to the description of inflectional morphology. This will require integrating detailed studies of affixation and of the structure and semantic compositionality of derived words.

Finally, G1 words are perhaps more difficult to process than G2 words. As [Hayes and Mouradian 81] put it:

> "Since novel words are by definition not in the known vocabulary, how can we distinguish them from misspelling?"

Most of the time (but not always) they will not be close enough to words in the dictionary for the system to make suggestions. The best one can hope for in this situation is to deduce from the context the maximum amount of information about the word, such as its role in the sentence. As for the ability to learn new vocabulary, this is beyond the capabilities of current artificial intelligence.

## 7.0 Acknowledgments

This work was conducted by the Computer-Assisted Translation group at the CCRIT. For their participation in the data collection we are greatly indebted to Pierre Isabelle, Elliott Macklovitch and Marie-Louise Hannan. We also wish to thank Marc Dymetman, Marie-Louise Hannan and Elliott Macklovitch for their helpful comments.

## 8.0 References

[Bourbeau, Pinard 86] Bourbeau, L. and Pinard, F., Dictionnaire micro-informatisé du Français (DMF), 1987, Progiciels Bourbeau-Pinard Inc., Montréal.

[Damerau 64] Damerau, F.J., A technique for computer detection and correction of spelling errors, Comm. ACM, 1964, 7, 3, pp. 171-176.

[Dinnematin and Sanz 90] Dinnematin, S., and Sanz, D., Sept correcteurs pour l'orthographe et la grammaire, Science et Vie Micro, 1990, pp. 118-130.

[Foster 91] Foster, G.F., Statistical lexical disambiguation, 1991, Master's thesis, McGill University, Montreal.

[Hayes and Mouradian 81] Hayes, P.J. and Mouradian, G.V., Flexible parsing, American Journal of Computational Linguistics, 81, 7, 4, pp. 232-242.

[Peterson 80] Peterson, J. L., Computer programs for detecting and correcting spelling errors, Comm. ACM, 1980, 23, pp. 676-687.

[Pollock 82] Pollock, J.J., Spelling error detection and correction by computer: some notes and a bibliography, Journal of Documentation, 1982, 38, 4, pp. 282-291.

[Pollock and Zamora 83] Pollock J.J. and Zamora, A., Collection and characterization of spelling errors in scientific and scholarly texts, J. Am. Soc. Inf. Sc., 1983, 34, 1, pp. 51-58.

[Shiaty 88] Shiaty, A.E. ed., Dictionnaire du français plus, 1988, CEC, Montréal.

[Srihari 85] Srihari, S.N., Computer text recognition and error correction, 1985, IEEE Computer Society Press, Silver Spring.

[Szanzer 69] Szanzer, A.J., Error-correcting methods in natural language processing, Information Processing, 1969, 68, 2.

[Veronis 88] Veronis, J., Correction of phonographic errors in natural language interfaces, Comm. ACM, 1988, pp. 101-115.

[Wagner and Fischer 74] Wagner, R.A. and Fischer, M. J., The string-to-string correction problem, JACM, 1974, 21, 1, pp. 168-178.