***1. Describe the difficulties of real-word errors. Limit your answer to 100 words.[1 point]***

Unlike non-word errors which could be detected and corrected by comparing to a list of correctly spelled words such as dictionaries, real-word errors are real words, namely they might be listed in that list. We then can’t solely rely on misspelling but more on the surrounding context of words, which is not a simple comparison taks, we might need to introduce language models to understand the context and then correct the errors based on rankings of all candidates.

***2. Examine the text below and list at least four different errors. For each, provide an error type, a likely cause, and a sequence of edit operations that correct the misspelling. [3 points]***

*“City of Glass” is a story about a lonely man named Quinn, whose a writer of detective stories. In the middle of the night he recieves a strange phonecall and some one on the other end asks him for help in a criminal vase. The person who called believes that Quinn is a detectiv named Paul Auster and wants to meet him. After the third call, Quinn decides to take Auster‘s identitz and goes to the meeting as a man he isn‘t. And and so he gets involved into a case that can‘t be more confusing and puzling.*

*(adapted from* [*https://schulzeug.at/englisch/summary/paul-auster-city-of-glass/*](https://schulzeug.at/englisch/summary/paul-auster-city-of-glass/)*)*

1. Spelling confusions: as in “recieves”, “detectiv”, “identitz”, “puzling”

Likely causes: not familiar with the spelling and typed them based on sound.

Edit operations:

* Insertion: recieves – rec**e**ieves
* Deletion: recei**e**ves - receives
* Or simply transposition: rec**ie**ves – rec**ei**ves

2. Typological errors: as in “phonecall”, “some one”

Likely causes: mis-sued the space bar

Edit operations:

* Insertion (Add a space): phonecall – phone call

3. Semantic error: “vase”

Likely causes: letter “v” and “c” are too close on the keyboard

Edit operations:

* Substitution: vase – case

4. Local syntactic error: “whose”, “And and”

Likely causes: “whose” might be a pure grammar mistake and should refer to Quinn, “And and” might happen when typed “and” first and then the computer autocorrects and adds an “And”.

Edit operations:

* Deletion: whose – who
* Insertion: who – who is
* Deletion: And and – And

***3. List all word uni-grams and bi-grams together with their frequencies for the following text [3 points]***

*a cat and a dog went to the beach, and a beautiful day it was!*

| Unigrams | Frequencies | Bigrams | Frequencies |
| --- | --- | --- | --- |
| a | 3 | a cat | 1 |
| cat | 1 | cat and | 1 |
| and | 2 | and a | 2 |
| dog | 1 | a dog | 1 |
| went | 1 | dog went | 1 |
| to | 1 | went to | 1 |
| the | 1 | to the | 1 |
| beach | 1 | the beach | 1 |
| beautical | 1 | beach and | 1 |
| day | 1 | a beautiful | 1 |
| it | 1 | beautiful day | 1 |
| was | 1 | day it | 1 |
| , | 1 | it was | 1 |
| ! | 1 | beach, | 1 |
|  |  | ,and | 1 |
|  |  | was! | 1 |

***4. Search the web for a linguistic corpus that you find interesting. Next, describe your choice in one or two paragraphs. Limit your answer to 350 words and be sure to include the following information: [3 points]***

* *Type of linguistic data (e.g., language, written or spoken, speech acts, pronunciation, etc.)*
* *Intended purpose of corpus*
* *Annotation types (global metadata, linguistic annotations)*
* *Method used for collection*
* *Data serialisation (e.g., file formats)*
* *Licensing information (e.g., freely available, or subscription required)*

With a primary purpose to represent a cross-section of British English in the late 20th century, the British National Corpus (BNC) is a comprehensive linguistic dataset containing 100 million words in both written and spoken English samples from diverse sources including materials such as newspapers, periodicals, academic books, fiction, and unscripted conversations. The corpus is monolingual, focusing on modern British English, and is synchronic, capturing the language of its time. It is a general corpus, incorporating various styles and genres, providing a full view of the language. The corpus is annotated according to the Text Encoding Initiative (TEI) guidelines, incorporating both global metadata and linguistic annotations such as part-of-speech tags generated by the CLAWS tagger. The data serialisation is based on the TEI standard, and the corpus is encoded in XML format.

The BNC's creation involved meticulous planning by the BNC Consortium, led by Oxford University Press, with contributions from major dictionary publishers, academic research centres, and governmental support. The corpus is freely available for research purposes.

**Reference**

BNC Consortium. (2007). *British National Corpus.* Retrieved from <http://www.natcorp.ox.ac.uk/>.