F. The lost tram

F1. The deviations in each text fragment are marked in bold and corrected:

(1)
The **tram (->train)** makes no stops; you sit **clown (->down)** and are served; there are no further intrusions, no **late-corners (->late-comers)**, no one hurrying to get off. The businessmen leaf through their financial reports, the lady with the hatbox is alone with her novel and her sirloin. Diners reading: you never see that on a plane. When the coast approaches **arid (->and)** dinner is over, everyone retires to his compartment to **he (->be)** transferred to the boat in peace, horizontally.
(Sunrise With Seamonsters, by Paul Theroux)

(2)
Usually, Howie could legitimately claim to have no **dear (->fear)** of any man or beast… Howie knew in his heart that it was **he (->the)** vulnerable positions he ended up in that scared him. He was used to operating from a position of strength, either real or projected. Now here he was, injured and alone, standing with **and (->an)** empty handgun in an open **filed (->field)**, while **hid (->his)** opponent or opponents **fried (->fired)** their weapon from behind solid cover.
(Rough Justice, by Mark Johnstone)

(3)
Two other factors **effect (->affect)** the body's temperature regulation: age and acclimatization. As we grow older, we **loose (->lose)** our ability to quickly regulate temperature… Very small children are also subject to heat disorders. **There (->Their)** small size allows them to take on heat much faster **then (->than)** adults. They also cannot indicate their thirst, **accept (->except)** through irritability. They are completely dependent upon adults to make certain they get enough fluids.
(Doctor in the House: Your Best Guide to Effective Medical Self-Care, by John Harbert)

F2. In the first text fragment, graphically similar letters or letter combinations are mixed up: **m<->in, m<->rn, ri<->n, cl<->d, h<->b.** This might have occurred if the text (probably messily printed or handwritten) had been interpreted by a computer (using an OCR, optical character recognition software) or (less probably) by a human who hadn't been paying attention to what he had been reading.

In the second text fragment, letters are skipped, added, rearranged or replaced by other letters (in the latter case, the pairs of letters corresponded to neighboring keys on a standard QWERTY keyboard: **d<->f, d<->s**). This most probably occurred when someone was typing too fast.

In the third text fragment, there are several lexical errors, when words with identical or very similar pronunciation are mixed up. This might have occurred if the person who had copied the text was quite bad in spelling, or maybe if the text was analyzed by a speech recognition system.

F3. Common spellchecking programs would not be of much help, since all wrong words are still English words (maybe the texts had already been through a spellcheck). To find at least a partial solution to fixing such deviations, one might create huge lists containing (1) common OCR mistakes (pairs of graphically similar words), (2) common misprints,

(3) commonly confused words. Some such lists already exist. Then, one could trace some (probably not all) mistakes using two alternative approaches. First, one could parse the texts using a natural language processing system, which might find some grammatical (mostly syntactical) mistakes. Constructing such systems is a very topical issue in modern computational linguistics, and a very complicated task. Second, one could verify all suspicious word combinations by searching them in a large text corpus, database, or simply in the web, and compare the number of hits to that of the alternative combination found in the lists. For example, a Google search yields some 8,690,000 results for **sit down**, and only 252 results for **sit clown** (probably most of them containing the same OCR error). This approach, however, only works for frequent word combinations and could accidentally result in wrong corrections for some rare, but not erroneous combinations. Therefore, the program should be an interactive one, marking potential mistakes and offering the user a variety of ways to correct them, but not attempting to correct them automatically.